BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Appellant(s):	Walke *et al.*	Group Art Unit: 1654
Application No.:	09/813,290	Examiner: B. D. Chism
Filed:	03/20/01	
Title:	Human Secreted Proteins and Polynucleotides Encoding the Same (As amended)	Attorney Docket No.: LEX-0151-USA

APPEAL BRIEF

Mail Stop Appeal Brief
Commissioner for Patents
Alexandria, VA 22313

TABLE OF CONTENTS

I. REAL PARTY IN INTEREST 1

II. RELATED APPEALS AND INTERFERENCES 1

III. STATUS OF THE CLAIMS 1-3

IV. STATUS OF THE AMENDMENTS 3-4

V. SUMMARY OF THE INVENTION 4

VI. ISSUES ON APPEAL 4

VII. GROUPING OF THE CLAIMS 5

VIII. ARGUMENT 5-26
- A. Do Claims 1-4, 11 and 12 Lack a Patentable Utility? 5-24
- B. Are Claims 1-4, 11 and 12 Unusable Due to a Lack of Patentable Utility? . 24-25
- C. Is Claim 11 indefinite? 25-26

IX. APPENDIX 27

X. CONCLUSION 28

TABLE OF AUTHORITIES

CASES

Amgen, Inc. v. Chugai Pharmaceutical Co., Ltd., 927 F.2d 1200, 18 USPQ2d 1016 (Fed. Cir. 1991) 7

Brooktree Corp. v. Advanced Micro Devices, Inc., 977 F.2d 1555, 1571, 24 USPQ2d 1401 (Fed. Cir. 1992) 5

Cross v. Iizuka, 753 F.2d 1040, 224 USPQ 739 (Fed. Cir. 1985) 5

Diamond vs. Chakrabarty, 447 U.S. 303, 206 USPQ 193 (U.S., 1980) 5

Envirotech Corp. v. Al George, Inc., 221 USPQ 473, 480 (Fed. Cir. 1984) 16

Hoffman v. Klaus, 9 USPQ2d 1657 (Bd. Pat. App. & Inter. 1988) 19

In re Angstadt and Griffin, 537 F.2d 498, 190 USPQ 214 (CCPA 1976) 7

In re Brana, 51 F.3d 1560, 34 USPQ2d 1436 (Fed. Cir. 1995) 6, 24

In re Fouche, 439 F.2d 1237, 1243, 169 USPQ 429, 434 (CCPA 1971) 24

In re Gay, 135 USPQ 311 (C.C.P.A. 1962) 11

In re Gottlieb, 328 F.2d 1016, 140 USPQ 665 (CCPA 1964) 19

In re Jolles, 628 F.2d 1322, 1326 n.11, 206 USPQ 885, 889 n.11 (CCPA 1980) 24

In re Langer, 503 F.2d 1380, 1391, 183 USPQ 288, 297 (CCPA, 1974) 18, 22

In re Marzocchi, 439 F.2d 220, 224, 169 USPQ 367, 370 (CCPA, 1971) 19, 22

In re Malachowski, 530 F.2d 1402, 189 USPQ 432 (CCPA 1976) 19

In re Wands, 858 F.2d 731, 8 USPQ 2d 1400 (Fed. Cir. 1988) 7

Juicy Whip Inc. v. Orange Bang Inc., 185 F.3d 1364, 51 USPQ2d 1700 (Fed. Cir. 1999) (citing *Brenner v. Manson*, 383 U.S. 519, 534 (1966)) 5

Raytheon Co. v. Roper Corp., 724 F.2d 951, 220 USPQ 592 (Fed. Cir. 1983) 19

State Street Bank & Trust Co. v. Signature Financial Group Inc., 149 F.3d 1368, 47 USPQ2d 1596, 1600 (Fed. Cir. 1998) 5

Carl Zeiss Stiftung v. Renishaw PLC, 20 USPQ2d 1101 (Fed. Cir. 1991) 16, 21

STATUTES

35 U.S.C. § 101 ..... 2, 4, 5-8, 10, 13, 16-19, 22-24

35 U.S.C. § 112 ..... 2, 5, 6, 7, 10, 13, 22-26

APPEAL BRIEF

Sir:

Appellants hereby submit an original and two copies of this Appeal Brief to the Board of Patent Appeals and Interferences ("the Board") in response to the Final Office Action mailed on May 20, 2003. The Notice of Appeal was timely submitted on September 22, 2003, and was received in the Patent and Trademark Office ("the Office") on September 26, 2003. This Appeal Brief is timely submitted in light of the concurrently filed Petition for an Extension of Time of four months to and including March 26, 2004 and authorization to deduct the fee as required under 37 C.F.R. § 1.17(a)(2) from Appellants' Representatives' deposit account. The Commissioner is also authorized to charge the fee for filing this Appeal Brief (**$165.00**), as required under 37 C.F.R. § 1.17(c), to Lexicon Genetics Incorporated Deposit Account No. 50-0892.

Appellants believe no fees in addition to the fee for filing the Appeal Brief and the fee for the extension of time are due in connection with this Appeal Brief. However, should any additional fees under 37 C.F.R. §§ 1.16 to 1.21 be required for any reason related to this communication, the Commissioner is authorized to charge any underpayment or credit any overpayment to Lexicon Genetics Incorporated Deposit Account No. 50-0892.

I. REAL PARTY IN INTEREST

The real party in interest is the Assignee, Lexicon Genetics Incorporated, 8800 Technology Forest Place, The Woodlands, Texas, 77381.

II. RELATED APPEALS AND INTERFERENCES

Appellants know of no related appeals or interferences.

III. STATUS OF THE CLAIMS

The present application was filed on March 20, 2001, claiming the benefit of U.S. Provisional

Application Numbers 60/190,638, 60/191,188, and 60/193,639 which were filed on March 20, 2000, March 22, 2000, and March 31, 2000, respectively, and included original claims 1-10. A Restriction and Election Requirement was issued by the Office on August 16, 2002, restricting the original claims into three separate and distinct inventions. In a response to the Restriction Requirement, submitted to the Office on September 16, 2001, Appellants elected without traverse the Group I invention (comprising original claims 1-4) for prosecution on the merits and as a result claims 5-10 were canceled without prejudice and disclaimer as being drawn to a non-elected inventions. In a First Official Action, issued on October 22, 2002 ("the First Action"), the Examiner objected to the title of the specification and rejected claims 1-4 under 35 U.S.C. § 112, first paragraph, allegedly due to a lack of written description. Claim 1 was also rejected under 35 U.S.C. § 112, second paragraph, as being allegedly indefinite for recitation of the phrase "sequence first disclosed in SEQ ID NO:1". Claim 2 was also rejected under 35 U.S.C. § 112, second paragraph, as being allegedly indefinite for recitation of the phrase "stringent conditions". In addition, claims 1-4 were also rejected under 35 U.S.C. § 101, due to the alleged lack of patentable utility, and under 35 U.S.C. § 112, first paragraph, as allegedly unusable by the skilled artisan due to the alleged lack of patentable utility. In Appellants' response to the First Official Action, submitted to the Office on February 12, 2003 ("response to the First Action"), Appellants amended the title of the specification, Claim 1 to further improve its clarity and added new claims 11-12 to more particularly point out and distinctly claim the present invention. A Second and Final Official Action, was issued on May 20, 2003 (the "Final Action"), in which it was stated that "The rejections and/or objections made in the prior office action, which are not explicitly stated below, in original or modified form are withdrawn" (page 2). Therefore objection to the title and rejections of claims 1 and 2 under 35 U.S.C. § 112, second paragraph, as being allegedly indefinite were withdrawn. The pending rejections of claims 1-4 under 35 U.S.C. § 101, and under 35 U.S.C. § 112, first paragraph, due to the alleged lack of patentable utility were maintained. Additionally, claim 11 was rejected under 35 U.S.C. § 112, second paragraph, as being allegedly indefinite for recitation of the allegedly indefinite recitation of the phrase "comprising [a] nucleic acid sequence of Claim 4." In a response to the Final Action, submitted on September 22, 2003 ("response to the Final Action"), an amendment to claim 11 was submitted and Appellants again addressed the outstanding rejections of claims

1-4 ,11 and12. A Notice of Appeal was also filed on September 22, 2003 and received by the U.S.P.T.O. on September 26, 2003. On December 5, 2003, Appellants received an atypical message from the Examiner that " that the current After Final was under consideration for allowable subject matter". This message was followed by a confirming interview summary (Paper No.20031201). As Appellants believed that the case contained allowable subject matter and based on the information in the Examiner's telephone message and the confirming Interview summary supporting this position, Appellants did not immediately submit an Appeal Brief. Following several unanswered telephone communications Appellants reached the Examiner during the week of February 23, 2004. At which time they were told that the Examiner did not have ready access to the case, but that he would investigate. Appellants left several more telephone messages during the week of March 15, 2004 and eventually received a return call in which the Examiner represented that he had the case and was addressing the issue. During the week of March 22, 2004 Appellants called the Examiner multiple times to determine the status of the case. On March 22, 2004 Appellants were told that the Examiner was prepared to discuss the case with his Supervisory Patent Examiner the next day. Appellants left additional messages and following a message on March 25, 2004, the Examiner left a return message that the case contained no allowable material. The Examiner alleged that he would mail out and Advisory Action shortly. Appellants note that at no time did the Examiner initiate contact to correct the misperception that the case contained allowable material. As Appellants had already lost potential patent term and accumulated costs in extension fees, and faced a 24 hour deadline for additional fees and because Appellants had no way of knowing what said Advisory Action might say, when and if it arrived, the present Appeal Brief and a 4 month extension of time has been filed. Thus, this Appeal Brief is based on Appellants last official written communication with the Office. A copy of the appealed claims (as of the Final Office Action) is included below in the Appendix (**Section IX**).

IV. STATUS OF THE AMENDMENTS

Appellants filed a response to the Final Office Action on September 22, 2003 that contained amendments to the claims. As at this time Appellants have received no Advisory Action entering these amendments into the case, Appellants must assume the these proposed amendments were not entered in

the case and are therefore outstanding.

V. SUMMARY OF THE INVENTION

The present invention relates to Appellants' discovery and identification of novel human polynucleotide and amino acid sequences that encode a novel human semaphorin protein (specification at or about page 2, lines 5-8 and 14-15; page 4, line 10 and page 17, lines 10 and 14-18). Semaphorins are a class of molecules with recognized function and utility having been implicated in mediating neural processes, cancer, and development. The semaphorin of the present invention was shown to be expressed in human fetal brain, brain, cerebellum, thymus, spleen, lymph node, kidney, uterus, adipose, esophagus, cervix, rectum, pericardium, and placenta (specification at or about page 4, lines 12-15) and those of skill in that art recognize that semaphorins are known to act to regulate the organization and fasciculation of nerves in the body (specification at or about page 2, lines 5-8). Thus the sequences of the present invention encode a molecule with specific, substantial and well-established function and utility. Additional uses described in the specification include assessing temporal and tissue specific gene expression patterns (at or about specification at page 7, line 23), particularly using a high throughput "chip" format (specification at page 6, line 29 through page 9), mapping the sequences to a specific region of a human chromosome and identifying protein encoding regions (specification at or about page 3, line 11), determining the genomic structure (specification at or about page 12, line 4), identifying verified intron/exon splice junctions (specification at or bout page 12, lines 5-10) and in diagnostic assays such as forensic analysis, human population biology and paternity determinations (see, for example, the specification at or about page 9, line 7; page 12, line 5 and page 18, line 11).

VI. ISSUES ON APPEAL

1. Do claims 1-4, 11 and 12 lack a patentable utility?
2. Are claims 1-4, 11 and 12 unusable by a skilled artisan due to a lack of patentable utility?
3. Is claim 11 indefiinite?

4. GROUPING OF THE CLAIMS

For the purposes of the outstanding rejections under 35 U.S.C. § 101 and 35 U.S.C. § 112, first paragraph, the claims will stand or fall together. The rejection of claim 11 under 35 U.S.C. § 112, first paragraph for allegedly being indefinite will stand and fall alone.

5. ARGUMENT

A. Do Claims 1-4, 11 and 12 Lack a Patentable Utility?

The Final Action first rejects claims 1-4, 11 and 12 under 35 U.S.C. § 101, as allegedly lacking a patentable utility due to not being supported by either a specific and substantial utility or a well-established utility. Appellants strongly disagree.

Appellants respectfully submit that the question of utility is a straightforward one as established by the courts. As set forth by the Federal Circuit, "(t)he threshold of utility is not high: An invention is 'useful' under section 101 if it is capable of providing some identifiable benefit." *Juicy Whip Inc. v. Orange Bang Inc.*, 51 USPQ2d 1700 (Fed. Cir. 1999) (citing *Brenner v. Manson*, 383 U.S. 519, 534 (1966)). Additionally, the Federal Circuit has stated that "(t)o violate § 101 the claimed device must be totally incapable of achieving a useful result." *Brooktree Corp. v. Advanced Micro Devices, Inc.*, 977 F.2d 1555, 1571 (Fed. Cir. 1992), emphasis added. *Cross v. Iizuka* (224 USPQ 739 (Fed. Cir. 1985); "*Cross*") states "any utility of the claimed compounds is sufficient to satisfy 35 U.S.C. § 101". *Cross* at 748, emphasis added. Indeed, the Federal Circuit recently emphatically confirmed that "anything under the sun that is made by man" is patentable *(State Street Bank & Trust Co. v. Signature Financial Group Inc.*, 47 USPQ2d 1596, 1600 (Fed. Cir. 1998), citing the U.S. Supreme Court's decision in *Diamond vs. Chakrabarty*, 206 USPQ 193 (S.Ct. 1980)).

The legal test for utility simply involves an assessment of whether those skilled in the art would find any of the utilities described for the invention to be credible or believable. According to the Examination Guidelines for the Utility Requirement, if the applicant has asserted that the claimed invention is useful for any particular purpose (i.e., it has a "specific and substantial utility") and the assertion would be considered credible by a person of ordinary skill in the art, the Examiner should not impose a rejection based on lack

of utility (66 Federal Register 1098, January 5, 2001).

In *In re Brana,* (34 USPQ2d 1436 (Fed. Cir. 1995), "*Brana*"), the Federal Circuit admonished the P.T.O. for confusing "the requirements under the law for obtaining a patent with the requirements for obtaining government approval to market a particular drug for human consumption". *Brana* at 1442. The Federal Circuit went on to state:

> At issue in this case is an important question of the legal constraints on patent office examination practice and policy. The question is, with regard to pharmaceutical inventions, what must the applicant provide regarding the practical utility or usefulness of the invention for which patent protection is sought. This is not a new issue; it is one which we would have thought had been settled by case law years ago.

Brana at 1439, emphasis added. The choice of the phrase "utility or usefulness" in the foregoing quotation is highly pertinent. The Federal Circuit is evidently using "utility" to refer to rejections under 35 U.S.C. § 101, and is using "usefulness" to refer to rejections under 35 U.S.C. § 112, first paragraph. This is made evident in the continuing text in *Brana*, which explains the correlation between 35 U.S.C. §§ 101 and 112, first paragraph. The Federal Circuit concluded:

> FDA approval, however, is not a prerequisite for finding a compound useful within the meaning of the patent laws. Usefulness in patent law, and in particular in the context of pharmaceutical inventions, necessarily includes the expectation of further research and development. The stage at which an invention in this field becomes useful is well before it is ready to be administered to humans. Were we to require Phase II testing in order to prove utility, the associated costs would prevent many companies from obtaining patent protection on promising new inventions, thereby eliminating an incentive to pursue, through research and development, potential cures in many crucial areas such as the treatment of cancer.

Brana at 1442-1443, citations omitted. In assessing the question of whether undue experimentation would be required in order to practice the claimed invention, the key term is "undue", not

"experimentation". *In re Angstadt and Griffin*, 190 USPQ 214 (C.C.P.A. 1976). The need for some experimentation does not render the claimed invention unpatentable. Indeed, a considerable amount of experimentation may be permissible if such experimentation is routinely practiced in the art. *In re Angstadt and Griffin, supra*; *Amgen, Inc. v. Chugai Pharmaceutical Co., Ltd.*, 18 USPQ2d 1016 (Fed. Cir. 1991). As a matter of law, it is well settled that a patent need not disclose what is well known in the art. *In re Wands*, 8 USPQ 2d 1400 (Fed. Cir. 1988).

Even under the newly installed utility guidelines, Appellants note that MPEP 2107 (II)(B)(1) states:

> (1) If the applicant has asserted that the claimed invention is useful for any particular practical purpose (i.e., it has a "specific and substantial utility") and the assertion would be considered credible by a person of ordinary skill in the art, do not impose a rejection based on lack of utility. (MPEP 2107 (II)(B)(1))

Presented in the First Official Action, and maintained in the Final Action, was the Examiner's position that the specification does not disclose a specific and substantial or well-established utility for the claimed invention. Appellants strongly disagree and note that in the specification (at or about page 2, lines 5-8 and 14-15; page 4, line 10 and page 17, lines 10 and 14-18) it was asserted that the sequences of the present invention encode a human semaphorin protein that is structurally similar to other known semaphorins. These statements in the specification assert that the sequences of the present invention and known semaphorins share a similarity in structure, a similarity in function and a similarity in biological function. This would be accepted by those of skill in the art, as it is generally recognized that there is a structure-function relationship. Thus clearly the sequences of the present invention have patentable utility and pending rejections under 35 U.S.C. § 101 and 35 U.S.C. § 112, first paragraph should be withdrawn.

First, as set forth in the response to the First Action, reiterated in the response to the Final Action, but never addressed in any of the Actions, Appellants would like to invite the Board's attention to the fact that a sequence, that is 99.872% identical over its entire length and which encompases 89.5

% (782/874 of the amino acids) of the full length of the described sequence (SEQ ID NO:3) is present in the leading scientific repository for biological sequence data (GenBank), and has been annotated by third party scientists wholly unaffiliated with Appellants as as encoding semaphorin sem2 [*Homo sapiens*] (GenBank accession no. BAA98132 alignment and information previously provided and as **Exhibit A**). Also as previously submitted was evidence in the form of a nucleic acid comparison between SEQ ID NO:3 and GenBank accession no. AB029496.1 (alignment and information previously provided and as **Exhibit B**), identified as *Homo sapiens* mRNA for semaphorin sem2. Thus clearly the identity between the sequences of the present invention and human semaphorin sem2 also exists at the nucleic acid level. Furthermore, as previously submitted in the response to the First Action is the results of a nucleic acid sequence comparison between SEQ ID NO:1 and SEQ ID NO:3 of the present invention, clearly indicating that SEQ ID NO:1 (see information previously provided and as **Exhibit C** comparing SEQ ID NOS: 3 and 1) identifies a longer isoform of the present invention, which is clearly encoded by the same genetic locus. Both the molecules described in SEQ ID NOS: 3 and 1 contain the recognized semaphorin signaling domain present in human semaphorin sem2 and the sema domains which are known to those of skill in the art to occur in semaphorins. Thus these molecules contain all the functional structure and domains required for them to function as a signaling semaphorin. Without doubt, those of skill in the art would readily recognize the sequences of the present invention as encoding a functioning human semaphorin.

Additionally, Appellants respectfully submit that human semaphorin are well known to those of skill in the art, semaphorins have recognized function and utility having been implicated in mediating neural processes and those of skill in that art recognize that semaphorins are known (as represented in the specification at or about page 2, lines 5-8) to act to regulate the organization and fasciculation of nerves in the body. Thus the sequences of the present invention encode a molecule with specific, substantial and well-established function and utility.

Clearly those of skill in the art would recognize the sequences of the present invention as encoding a human semaphorin. As evidenced by the review article entitled "Molecular Mechanisms of Axonal Guidance" from the prestigious journal Science (**298**:1959-1964, 2002 and erratum; document

previously provided and as **Exhibit D**), semaphorins are well known to those of skill in the art as soluble and membrane-bound proteins that act as chemorepulsive factors in neuronal development, thereby playing a crucial role in axon guidance. Semaphorins, such as the one described in the present invention, provide guidance for neuronal growth. In the second paragraph of section 5.1 or the specification as filed, it is stated that "Because of their role in neural development, semaphorins have been subject to considerable scientific scrutiny. For example, U.S. Patents Nos. 5,981,222 and 5,935,865, both of which are herein incorporated by reference, describe other semaphorins as well as applications, utilities". Therefore, clearly, there can be no question that Appellants' asserted identity and utility for the described sequences a semaphorin is "credible." In addition, those of skill in the art in the biomedical and pharmaceutical industry would readily recognize the utility for semaphorins and their application to medical conditions requiring nerve regeneration. For example, the regeneration and repair of nerve tissue following the surgical attachment of severed limbs or the resection of diseased tissue, as well as nerve repair following a stroke. The specification details tissues in which these sequences are expressed (human fetal brain, brain, cerebellum and others at or about page 4, lines 12-15) and disease associations, both of which are consistent with the evidence provided and asserted utility.

Thus Appellants have provided evidence that the sequences of the present invention which were asserted in the specification to encode novel human semaphorin proteins do indeed encode the human semaphorin proteins (specifically longerisoforms of sem2). This evidence includes sequence identity, tissue expression, disease association and, below, genetic mapping to the same loci. Therefore, clearly, there can be no question that Appellants' asserted utility for the described sequences is "credible" and that those of skill in the art would recognize that the sequences of the present invention encode a semaphorin protein, more particularlyisoforms of sem2 and has all the recognized uses thereof. In contrast, the Examiner has provided no evidence of record indicating that those of skill in the art would not recognize the sequences of the present invention encode semaphorin proteins. As such, the scientific evidence clearly establishes that Appellants have described an invention having a specific, substantial and well-established utility and whose utility is in full compliance with the provisions

of 35 U.S.C. § 101, and the Examiner's rejection should be overturned.

Furthermore, Appellants respectfully submit that the Examiner's position, in light of the evidence provided, runs contrary to Example 10 of the PTO's Revised Interim Utility Guidelines Training Materials (pages 53-55), which establishes that a rejection under 35 U.S.C. § 101 as allegedly lacking a patentable utility and under 35 U.S.C. § 112, first paragraph as allegedly unusable by the skilled artisan due to the alleged lack of patentable utility, is not proper when there is no reason to doubt the asserted utility of a full length sequence (such as the presently claimed sequence) that has a high degree of similarity to a protein having a known function. In the Analysis portion of Example 10 it states that "Based on applicant's disclosure and the results of the PTO search, there is no reason to doubt the assertion that SEQ ID NO:2 encodes a DNA ligase. Further DNA ligases have a well-established use in the molecular biology art based on this class of proteins ability to ligate DNA. .....Note that if there is a well-established utility already associated with the claimed invention, the utility need not be asserted in the specification as filed..... Thus the conclusion reached from this analysis is that a 35 U.S.C. § 101 and a 35 U.S.C. § 112 first paragraph, utility rejection should not be made."

In the present case, clearly the evidence supports Appellants' assertions that the sequences of the present invention encode human semaphorin proteins (specifically isoforms of sem2), a class of proteins for which there is a well established utility that is recognized by those of skill in the art and a specific semaphorin whose function is known to those of skill in the art. Thus the present case is identical to that presented in Example 10 of the Revised Interim Utility Guidelines Training Materials (pages 53-55). In the present case it is clear that the sequences of the present invention encode human semaphorin proteins (specifically isoforms of sem2). The Examiner dismisses Appellants' continued assertions and the evidence provided that the protein of the present invention are semaphorins (specifically isoforms of sem2) and that the function of semaphorins as a class of proteins is well known to those of skill in the the art. Thus, according to the guidelines the conclusion reached from this analysis is that a 35 U.S.C. § 101 and a 35 U.S.C. § 112 first paragraph, utility rejection should not have been made. Thus the rejection of the presently claimed invention under a 35 U.S.C. § 101 and a 35 U.S.C. § 112 first paragraph utility rejection should be overruled.

The First Action (and maintained in the Final Action) takes issue with the fact that the specification discloses no data for any activity of the present invention and that there are no working examples, indicating a need for such information is misplaced. It has long been established that "there is no statutory requirement for the disclosure of a specific example". *In re Gay*, 135 USPQ 311 (C.C.P.A. 1962). The Actions also assume the position that structural homology cannot be accepted in the absence of supporting evidence, because the relevant literature acknowledges that function cannot be based solely on structural similarity to a protein found in the sequence database. In support of this position the Final Action cites Bork (Genome Research *10*:398-400, 2000) as supporting the proposition that prediction of protein function from homology information is somewhat unpredictable. It is of interest that in his "analysis" Bork often uses citations to many of his own previous publications, an interesting approach. ' My position is supported by my previous disclosures of my position.' If Bork's position is supported by others of skill in the art, one would expect that he would reference them rather than himself to provide support for his statements. Given that the standard with regard to obtaining U.S. patents is those of skill in the art, this observation casts doubt on the broad applicability of Bork's position. It should also be noted that in Table 1, on page 399, in which selected examples of prediction accuracy are presented, that the reported accuracy of the methods which Appellants have employed are, in fact, very high. While nowhere in Bork is there a comparison of the prediction accuracy based on the percentage homology between two proteins or two classes of proteins, " Homology (several methods)" is assigned an accuracy rate of 98% and "Functional features by homology" is assigned an accuracy rate of 90%. Given that these figures were obtained based on what is at least a 4 year old analysis, these high levels of accuracy would appear to support rather than refute Appellants' assertions in the present case. Additionally Bork even states (on page 400, second column, line 17) that " However, there is still no doubt that sequence analysis is extremely powerful". In summary, it is clear that it is not Bork's intention to refute the value of sequence analysis but rather he is indicating that there is room for improvement .

In summary a careful reading of the cited "relevant literature" does not in fact support the concept that function cannot be based on sequence and structural similarity, in contrast many of the

examples actually support the use of such methodologies while identifying several areas in which caution should be exercised. These inaccuracies and potential pitfalls can be overcome by a more careful analysis by those of skill in the art. Automatic methods of sequence homology identification was only the staring point for consideration the sequences of the present invention underwent careful analysis by a series of individuals of skill in the art, many highly qualified (experienced B.S. and 3 Ph.D. level scientists).

Furthermore, this articles is just an example of the few contrarian articles that the PTO has repeatedly attempted to use to deny the utility of nucleic acid sequences based on a small number of publications that call into doubt prediction of protein function from homology information and the usefulness of bioinformatic predictions. While there may not be a 100% consensus within the scientific community regarding prediction of protein function from homology information, this is not unusual nor is it indicative of a general lack of consensus. A few rare exceptions do not a rule make.

The position that bioinformatic information is recognized to be of value by those of skill in the art is supported by the results of a recent search of the NCBI-NLM-NIH public scientific database "PubMed" using the term "bioinformatics" which resulted in 5,548 different scientific publications (these will not be provided to avoid burdening the USPTOs scanning group). If bioinformatic information is not useful in predicting protein function from structural homology information, why are so many publications reporting the results of its use? Clearly this suggest that those of skill in the art do recognize bioinformatic data as useful and valid.

A second form of evidence supporting the position that bioinformatic information is recognized to be of value by those of skill in the art is the fact that many scientists, corporations and institutions elect to allocate significant proportions of their limited resources for access to private bioinformatic systems and databases. Thus, it would appear obvious that those of skill in the art value and accept the results of bioinformatic analysis for they are willing to pay dearly for access to such information.

A third, and perhaps most persuasive, form of evidence supporting the position that bioinformatic information is recognized to be of value by those of skill in the art is the issuance of multiple US patents regarding bioinformatic prediction and methods for doing the same (see for

example, U.S. Patent Nos. 6,229,911, 6,567,540, 6,615,141, 6,631,331, 6,651,008, 6,677,114, **Exhibits E-J**; copies of issued U.S. Patents not provided pursuant to current United States Patent and Trademark Office policy). Of particular interest might be U.S. Patent No. 6,466,874 (**Exhibit K**; copies of issued U.S. Patents not provided pursuant to current United States Patent and Trademark Office policy). one of whose claims reads on "A method of identifying proteins as functionally linked, the method comprising comparing sequences to find homologous functional domains." Why would a U.S. Patent have issued on a method of carrying out an analysis that is without utility, because it is not accepted by those of skill in the art as a credible method of predicting function from structural homology information? This evidence convincingly indicates that even the USPTO recognizes the utility of bioinformatic prediction.

Appellants respectfully point out that, as discussed above, the legal test for utility simply involves an assessment of whether those skilled in the art would find any of the utilities described for the invention to be believable. Appellants submit that the overwhelming majority of those of skill in the relevant art would believe prediction of protein function from homology information and the usefulness of bioinformatic predictions to be powerful and useful tools. Clearly the several forms of evidence presented, and certainly the issuance of U.S. Patents suggest that those of skill in the art recognize the utility of bioinformatic analysis and its credibility in assessing structure function relationships. Thus the vast majority of those of skill in the art would believe that Appellants' sequence encodes a human semaphorin proteins (specifically isoforms of sem2), a molecule of specific, substantial and well-established utility and thus rejection of the presently claimed invention under a 35 U.S.C. § 101 and a 35 U.S.C. § 112 first paragraph should be overruled.

In addition to those utilities presented above, a still further example of utility of the present sequences is their use in diagnostic assays such as those associated with identification of paternity and forensic analysis, among others (see, for example, the specification at or about page 9, line 7; page 12, line 5 and page 18, line 11). The sequences of the present invention have particular utility as the application as filed contained an identified polymorphism (at or about page 17, line 10-13). This results in a translationally silent A-to-G transition at, for example, the position corresponding to nucleotide

2106 of SEQ ID NO:1.

Naturally occurring genetic polymorphisms such as those described in the present specification are both the basis of, and critical to, *inter alia*, forensic genetic analysis and genetic analysis intended to resolve issues of identity and paternity. Therefore, Appellants find this position difficult to comprehend, given that the results of identity and paternal analysis often have great emotional and substantial economic impact. This does not sound like a throw away utility, rather it sounds like a very substantial and real world utility. What could be more substantial and real world than the loss of an individual's freedom through incarceration and in some cases even the loss of life through execution? Yet forensic analysis based on identified polymorphisms is often used to convict or acquit in many cases. Both paternal and forensic genetic analysis is based on the use of identified polymorphisms. This is a well known and generally accepted by those of skill in the art, who would readily recognize the utility and value of any identified polymorphism. Without identified polymorphisms, one would not be able to carry out such forensic or paternal analyses. The present application has identified just such essential polymorphisms within the sequences of the present invention which identify human semaphorin proteins (specifically isoforms of sem2), a molecule of well-established utility.

Such polymorphisms are the basis for forensic analysis, paternity identification and population biology studies, which are undoubtedly "real world" utilities and thus the present sequences must in themselves be useful. In and of themselves each of these polymorphisms, including the silent ones, has significant and specific utility, the specificity of this utility is only amplified by the presence of so many polymorphisms that can arise in various combinations. It is also important to note that the presence of more useful polymorphic markers for such analysis would not mean that the present sequences lack utility.

Appellants respectfully point out that those of skill in the art would readily recognize that the presently described polymorphisms, exactly as they were described in the specification as originally filed, are useful in forensic analysis, population biology and paternity analysis to specifically identify individual members of the human population based on the presence or absence of the described polymorphism. Simply because the use of these polymorphic markers will necessarily provide

additional information on the percentage of particular subpopulations that contain one or more of these polymorphic markers does not mean that "additional research" is needed in order for these markers as they are presently described in the instant specification to be of use to forensic science. Without further experimentation those of skill in the art would recognize the utility of the identified polymorphisms and how the asserted markers can distinguish 50% of the population in the worst case scenario. Thus the presence or the absence of a particular specific polymorphism is sufficient for use in the proposed utilities. Appellants provide the following detailed explanation. Those of skill in the art would recognize that in the worst case, least useful situation, a marker would be present in half of a population and absent from the other half. Therefore the probability of an individual having such a marker would be 1 in 2 or 50%. Using the forensic analysis scenario for example, the analysis will have removed 50% of the possible suspects from the list, as either the suspect has the identified polymorphism or not. However, if a polymorphism were present in only say 10% of the population, the probability of an individual having such a polymorphic marker would be 1 in 10 (10%) and 90% of suspects could be eliminated from investigation or prosecution based on the presence or absence of the polymorphism. Clearly eliminating 90% of the suspects is better than eliminating 50% of the suspects. That said, eliminating 50% or half of the suspects on a list is without question very useful to any investigator. To reiterate, using the polymorphic markers as described in the specification as originally field will definitely distinguish members of a population from one another. In the worst case scenario, each of these markers are useful to distinguish 50% of the population (in other words, the marker being present in half of the population). The ability to eliminate 50% of the population from a forensic analysis clearly is a real world, practical utility. Therefore, any allegation that the use of the presently described polymorphic markers is only potentially useful would be completely without merit, and would not support the alleged lack of utility.

The Examiner's assumption appears to be that since any human nucleic acid sequence that contains a naturally occurring polymorphism can be used in forensic analysis, in human paternity determinations or human population migration determinations, such utilities are generic and therefore lack substantial and specific utility. First, Appellants submit that until a specific polymorphic marker is

actually described it has very limited utility in forensic analysis. Put another way, simply because there is a possibility, even a significant likelihood, that a particular nucleic acid sequence will contain a polymorphism and thus be useful in forensic analysis, until such a specific polymorphism is actually identified and described, such a likelihood is meaningless. The present case contains identified polymorphisms that occur in human semaphorin proteins (specifically isoforms of sem2). The Examiner is perhaps attempting to use the information presented for the first time by Appellants in the instant specification as hindsight verification that the presently claimed sequence would be expected to have polymorphic markers. Such a hindsight analysis based on Appellants' discovery would not be proper.

Alternatively, the assumption that since any sequence containing a naturally occurring polymorphism can be used such utilities are generic and therefore lack substantial and specific utility may represent a confusion between the requirement for a specific utility, which is the proper standard for utility under 35 U.S.C. § 101, with a requirement for a unique utility. The relevant case law cited by Appellants makes it abundantly clear that the presence of other or even more useful polymorphic markers for forensic analysis does **not** mean that the present sequences lack a specific utility. As clearly stated by the Federal Circuit in *Carl Zeiss Stiftung v. Renishaw PLC*, 20 USPQ2d 1101 (Fed. Cir. 1991; "*Carl Zeiss*"):

> An invention need not be the best or only way to accomplish a certain result, and it need only be useful to some extent and in certain applications: "[T]he fact that an invention has only limited utility and is only operable in certain applications is not grounds for finding a lack of utility." *Envirotech Corp. v. Al George, Inc.*, 221 USPQ 473, 480 (Fed. Cir. 1984)

Importantly, the holding in the *Carl Zeiss* case is mandatory legal authority that essentially controls the outcome of the present appeal. This case, and particularly the cited quote, directly rebuts any such argument. Furthermore, the requirement for a unique utility is clearly not the standard adopted by the Patent and Trademark Office. If every invention were required to have a unique utility, the Patent and Trademark Office would no longer be issuing patents on batteries, automobile tires, golf balls, golf clubs, and treatments for a variety of human diseases, such as cancer and bacterial or viral infections, just to name a few particular examples, because examples of each of these have already been

described and patented. All batteries have the exact same utility - specifically, to provide power. All automobile tires have the exact same utility - specifically, for use on automobiles. All golf balls and golf clubs have the exact same utility - specifically, use in the game of golf. All cancer treatments have the exact same utility - specifically, to treat cancer. All anti-infectious agents have the exact same broader utility - specifically, to treat infections. However, only the briefest perusal of virtually any issue of the Official Gazette provides numerous examples of patents being granted on each of the above compositions every week. Furthermore, if a composition needed to be unique to be patented, the entire class and subclass system would be an effort in futility, as the class and subclass system serves solely to group such common inventions, which would not be required if each invention needed to have a unique utility. Thus, the present sequence clearly meets the requirements of 35 U.S.C. § 101.

In Addition, the First and Final Actions discount Appellants' assertion regarding the use of the presently claimed polynucleotides on DNA gene chips, based on the position that such a use would allegedly be generic. Further, these Actions seem to require Appellants to identify the biological role of the nucleic acid or function of the protein encoded by the presently claimed polynucleotides before the present sequences can be used in gene chip applications that meet the requirements of § 101. Appellants respectfully point out that knowledge of the exact function or role of the presently claimed sequence is not required to track expression patterns using a DNA chip. As set forth in Appellants' First Response, given the widespread utility of such "gene chip" methods using *public domain* gene sequence information, there can be little doubt that the use of the presently described *novel* sequences would have great utility in such DNA chip applications. Even though not a requirement for use of a sequence on a DNA chip, clearly, the claimed sequences which encode human semaphorin proteins (specificallyisoforms of sem2), a molecule of recognized function that is believed to play a role in human disease, provide a specific marker of the gene encoding this protein and provide a unique identifier of the corresponding gene in the human genome. Such specific markers are targets for discovering drugs that are associated with human semaphorins known to act to regulate the organization and fasciculation of nerves in the body and involved in human neural processes, stroke, cancer, and development, among others. Thus, those skilled in the art would instantly recognize that the present

nucleotide sequence would be an ideal, novel candidate for assessing gene expression using, for example, DNA chips, as the specification details at least on or about page 6, line1 through page 8. Such "DNA chips" clearly have utility, as evidenced by hundreds of issued U.S. Patents, exemplified by U.S. Patent Nos. 5,445,934, 5,556,752, 5,744,305, 5,837,832, 6,156,501 and 6,261,776 (**Exhibits L-Q**; copies of issued U.S. Patents not provided pursuant to current United States Patent and Trademark Office policy).

The Board is further requested to consider that, given the huge expense of the drug discovery process, even negative information has great "real world" practical utility. Knowing that a given gene is not expressed in medically relevant tissue provides an informative finding of great value to industry by allowing for the more efficient deployment of expensive drug discovery resources. Such practical considerations are equally applicable to the scientific community in general, in that time and resources are not wasted chasing what are essentially scientific dead-ends (from the perspective of medical relevance). Clearly, compositions that enhance the utility of such DNA gene chips, such as the presently claimed sequences human semaphorins (variants of sem2) associated with human disease, must in themselves be useful. Moreover, the presently described sequences which sequences human semaphorins (variants of sem2) provide uniquely specific sequence resources for identifying and quantifying full length transcripts that were encoded by the corresponding human genomic locus. Accordingly, there can be no question that the described sequences provide an exquisitely specific utility for analyzing gene expression. Apparently the Examiner sees no public benefit in drugs or diagnostic assays directed at human disease, such as those involving neural processes like stroke, cancer, and developmental abnormalities.

Additionally, only a small percentage of the genome (2-4%) actually encodes exons, which in turn encode amino acid sequences. Thus, not all human genomic DNA sequences are useful in such gene chip applications. This further discounts the Examiner's position that such uses are "generic". The present claims clearly meet the requirements of 35 U.S.C. § 101. It has been clearly established that a statement of utility in a specification must be accepted absent reasons why one skilled in the art would have reason to doubt the objective truth of such statement. *In re Langer*, 503 F.2d 1380,

1391, 183 USPQ 288, 297 (CCPA, 1974); *In re Marzocchi*, 439 F.2d 220, 224, 169 USPQ 367, 370 (CCPA, 1971).

Evidence of the "real world" substantial utility of the present invention is further provided by the fact that there is an entire industry based on the use of gene sequences or fragments thereof in a gene chip format. Perhaps the most notable gene chip company is Affymetrix. However, there are many companies which have, at one time or another, concentrated on the use of gene sequences or fragments, in gene chip and non-gene chip formats, for example: Gene Logic, ABI-Perkin-Elmer, HySeq and Incyte. In addition, one such company, Rosetta Inpharmatics, was viewed to have such "real world" value that it was acquired by large pharmaceutical company, Merck & Co., for substantial sums of money (net equity value of the transaction was $620 million). The "real world" substantial industrial utility of gene sequences or fragments would, therefore, appear to be widespread and well established. Clearly, persons of skill in the art, as well as venture capitalists and investors, readily recognize the utility, both scientific and commercial, of genomic data in general, and specifically human genomic data. Billions of dollars have been invested in the human genome project, resulting in useful genomic data (see, *e.g.*, Venter *et al.*, 2001, Science *291*:1304; **Exhibit R**). The results have been a stunning success as the utility of human genomic data has been widely recognized as a great gift to humanity (see, *e.g.*, Jasny and Kennedy, 2001, Science *291*:1153; **Exhibit S**). Clearly, the usefulness of human genomic data, such as the presently claimed nucleic acid molecules, is substantial and credible (worthy of billions of dollars and the creation of numerous companies focused on such information) and well-established (the utility of human genomic information has been clearly understood for many years).

Further evidence of utility of the presently claimed polynucleotide, although only one is needed to meet the requirements of 35 U.S.C. § 101 (*Raytheon v. Roper*, 220 USPQ 592 (Fed. Cir. 1983); *In re Gottlieb*, 140 USPQ 665 (CCPA 1964); *In re Malachowski*, 189 USPQ 432 (CCPA 1976); *Hoffman v. Klaus*, 9 USPQ2d 1657 (Bd. Pat. App. & Inter. 1988)), is the specific utility the present nucleotide sequence has in determining the genomic structure of the corresponding human chromosome (specification at or about page 12, line 4), for example mapping the protein encoding regions as described in the specification (specification at or about page 3, line 11 and page 12, lines 5-10) and as

evidenced in the response to the Final Action and reiterated below. Clearly, the present polynucleotide provides exquisite specificity in localizing the specific region of the human chromosome containing the gene encoding the given polynucleotide, a utility not shared by virtually any other nucleic acid sequence. In fact, it is this specificity that makes this particular sequence so useful. Early gene mapping techniques relied on methods such as Giemsa staining to identify regions of chromosomes. However, such techniques produced genetic maps with a resolution of only 5 to 10 megabases, far too low to be of much help in identifying specific genes involved in disease. The skilled artisan readily appreciates the significant benefit afforded by markers that map a specific locus of the human genome, such as the present nucleic acid sequence.

Only a minor percentage of the genome actually encodes exons, which in turn encode amino acid sequences. The presently claimed polynucleotide sequence provides biologically validated empirical data (*e.g.*, showing which sequences are transcribed, spliced, and polyadenylated) that *specifically* defines that portion of the corresponding genomic locus that actually encodes exon sequence. Equally significant is that the claimed polynucleotide sequence defines how the encoded exons are actually spliced together to produce an active transcript (*i.e.*, the described sequences are useful for functionally defining exon splice-junctions). The Appellants respectfully submit that the practical scientific value of expressed, spliced, and polyadenylated mRNA sequences is readily apparent to those skilled in the relevant biological and biochemical arts. For further evidence supporting the Appellants' position, the Board is requested to review, for example, section 3 of Venter *et al.* (*supra at* pp. 1317-1321, including Fig. 11 at pp.1324-1325), which demonstrates the significance of expressed sequence information in the structural analysis of genomic data. The presently claimed polynucleotide sequence defines a biologically validated sequence that provides a unique and specific resource for mapping the genome essentially as described in the Venter *et al.* article.

While it is clear that the present nucleotide sequences have <u>specific</u> utility in determining genomic structure, mapping of the corresponding human chromosome, and determining protein encoding regions as was described in the specification, discussed in Appellants' response to the First Action and both discussed and evidenced in Appellants' response to the Final actions is reiterated here.

Evidence supporting Appellants' assertions of the specific utility of the sequences of the present invention in localizing the specific region of the human chromosome and identification of functionally active intron/exon splice junctions is the information provided as **Exhibit T.** This is the result of overlaying the sequence of SEQ ID NO:1 of the present invention and the identified human genomic sequence. By doing this, one is able to identify the portions of the genome that encode the present invention. As these regions of the genome are non-contiguous, this is indicative of individual exons. The results of such an analysis indicate that the sequence of the present invention is the result of a 16 exon gene contained within the BAC clone AC006208.3. Clearly as the gene of the present invention is encoded by 16 non-contiguous exons on chromosome 3, one would not have been able to deduce the sequence that encodes the molecules of the present invention without knowing the specific sequence. Clearly, the present polynucleotide provides exquisite specificity in localizing the specific region of human chromosome 3 that contains the gene encoding the given polynucleotide, a utility not shared by virtually any other nucleic acid sequences. The sequences of the present invention provide that necessary specific prior knowledge. In fact, it is this specificity that makes this particular sequence so useful.

Additionally, it should be noted that the gene encoding BAA98132, **Exhibit A**, identified as *Homo sapiens* semaphorin protein sem2, also maps within the same region of human chromosome 3 (essentially position 3p31.31). Thus in addition to providing direct evidence of the utility of the sequences of the present invention in chromosome mapping, this evidence further supports Appellant's assertion that the sequences of the present invention encode isoforms of *Homo sapiens* semaphorin protein sem2.

The Examiner's repeated position that this utility, like the use of these specific sequences on DNA chips or the described polymorphisms in forensic analysis, is that since other molecules can be used to map the human chromosome or on DNA chips or in forensic analysis, these utilities are not specific or substantial. As described previously above, Appellants once again point out that these arguments are completely rebuffed by the Federal Circuit's holding in *Carl Zeiss*, *supra* ("[A]n invention need not be the best or only way to accomplish a certain result"). Furthermore, the argument

that just because there are other objects having the same utility, that utility has been rendered generic and therefore invalid begs the question, previously presented, that don't all golf balls and tires have the same utility of other golf balls or tires, i.e. they can be used as golf balls or tires respectively and yet these items are readily considered to have patentable utility.

It has been clearly established that a statement of utility in a specification must be accepted absent reasons why one skilled in the art would have reason to doubt the objective truth of such statement. *In re Langer*, 503 F.2d 1380, 1391, 183 USPQ 288, 297 (CCPA, 1974; "*Langer*"); *In re Marzocchi*, 439 F.2d 220, 224, 169 USPQ 367, 370 (CCPA, 1971). As clearly set forth in *Langer*:

> As a matter of Patent Office practice, a specification which contains a disclosure of utility which corresponds in scope to the subject matter sought to be patented must be taken as sufficient to satisfy the utility requirement of § 101 for the entire claimed subject matter unless there is a reason for one skilled in the art to question the objective truth of the statement of utility or its scope.

Langer at 297, emphasis in original. As set forth in the MPEP, "Office personnel must provide evidence sufficient to show that the statement of asserted utility would be considered 'false' by a person of ordinary skill in the art" (MPEP, Eighth Edition at 2100-40, emphasis added).

In the present case, Appellants have provided multiple forms of evidence supporting their assertion that the sequences of the present invention encode human semaphorin proteins (isoforms of sem2), molecules with specific, substantial and well-established utility. In contrast, the Examiner has failed to provide evidence that the asserted utilities would be considered 'false' by a person of ordinary skill in the art and therefore has failed to provide support for the pending utility rejections, as required by the Utility Guidelines and the law. Thus clearly the rejection of the presently claimed invention under a 35 U.S.C. § 101 and a 35 U.S.C. § 112 first paragraph utility rejection was improper and should be overruled.

Finally, with full recognition of the fact that all patent applications are examined on their own merits and that the prosecution of one patent does not effect the prosecution of another patent, *In re Wertheim,* 541 F.2d 257, 264, 191 USPQ 90, 97 (CCPA 1976), however the issue at hand in one of whether the fact that patents have issued recognizing the utility of a class of molecules does this confers a statutory precedent of patentability to a broad class of compositions. Thus, there remains a lingering issue regarding due process and equitable treatment under the law. While Appellants are well aware of

the new Utility Guidelines set forth by the USPTO, Appellants respectfully point out that the current rules and regulations regarding the examination of patent applications is and always has been the patent laws as set forth in 35 U.S.C. and the patent rules as set forth in 37 C.F.R., not the Manual of Patent Examination Procedure or particular guidelines for patent examination set forth by the USPTO. Furthermore, it is the job of the judiciary, not the USPTO, to interpret these laws and rules. Appellants are unaware of any significant recent changes in either 35 U.S.C. § 101, or in the interpretation of 35 U.S.C. § 101 by the Supreme Court or the Federal Circuit that is in keeping with the new Utility Guidelines set forth by the USPTO. This is underscored by numerous patents that have been issued over the years that claim nucleic acid fragments that do not comply with the new Utility Guidelines. As examples of such issued U.S. Patents, the Examiner is invited to review U.S. Patent Nos. 5,817,479, 5,654,173, and 5,552,281 (each of which claims short polynucleotides; **Exhibits U-W**; copies of issued U.S. Patents not provided pursuant to current United States Patent and Trademark Office policy), and recently issued U.S. Patent No. 6,340,583 (which includes no working examples; **Exhibit X**; copies of issued U.S. Patents not provided pursuant to current United States Patent and Trademark Office policy), none of which contain examples of the "real-world" utilities that the Examiner appears to desire. Given the rapid pace of development in the biotechnology arts, it is difficult for the Appellants to understand how an invention fully disclosed and free of prior art at the time the present application was filed, could somehow retain *less* utility and be *less* enabled than inventions in the cited issued U.S. patents (which were filed during a time when the level of skill in the art was clearly lower). Simply put, Appellants' invention is *more* enabled and retains *at least as much* utility as the inventions described in the claims of the U.S. patents of record. As issued U.S. Patents are presumed to meet all of the requirements for patentability, including 35 U.S.C. §§ 101 and 112, first paragraph, Appellants submit that the present polynucleotides must also meet the requirements of 35 U.S.C. § 101. While Appellants agree that each application is examined on its own merits, Appellants are unaware of any changes to 35 U.S.C. § 101, or in the interpretation of 35 U.S.C. § 101 by the Supreme Court or the Federal Circuit, since the issuance of these patents that render the subject matter claimed in these patents, which is similar to the subject matter in question in the present application, as suddenly non-

statutory or failing to meet the requirements of 35 U.S.C. § 101. Thus, holding Appellants' invention to a different standard of utility is inconsistent and inequitable, such a judgement being arbitrary and capricious, a violation of due process and equal protection under the law, cannot be maintained.

Thus in summary, Appellants' application described novel nucleic and amino acid sequences that encode human semaphorin proteins (isoforms of sem2), molecules with specific, substantial and well-established utility. Semaphorins have, as stated in the specification recognized associations with human development and disease. Furthermore, the application also described the tissue specific expression pattern and a naturally occurring polymorphism that occurs within the sequences of the present invention which provide additional utility. The present situation directly tracks Example 10 of the Revised Interim Utility Guidelines Training Materials (pages 53-55), which establishes that a rejection under 35 U.S.C. § 101 as allegedly lacking a patentable utility and under 35 U.S.C. § 112, first paragraph as allegedly unusable by the skilled artisan due to the alleged lack of patentable utility, is not proper when the full length sequence of the invention encodes a protein that has a well known function. Therefore, Appellants submit that as the presently claimed sequences have been shown to have a substantial, specific, credible and well-established utility, the rejection of the claims under 35 U.S.C. § 101 and 35 U.S.C. § 112 first paragraph was improper. Thus, Appellants respectfully submit that the utility rejection of the pending claims under 35 U.S.C. § 101 and 35 U.S.C. § 112 first paragraph must be overruled.

B. Are Claims 1-4, 11 and 12 Unusable Due to a Lack of Patentable Utility?

The Final Action next rejects claims 1-4, 11 and 12 under 35 U.S.C. § 112, first paragraph, since allegedly one skilled in the art would not know how to use the invention, as the invention allegedly is not supported by either a clear asserted utility or a well-established utility.

The arguments detailed above in **Section VIII(A)** concerning the utility of the presently claimed sequences are incorporated herein by reference. As the Federal Circuit and its predecessor have determined that the utility requirement of Section 101 and the how to use requirement of Section 112, first paragraph, have the same basis, specifically the disclosure of a credible utility (*In re Brana*, *supra*; *In re Jolles*, 628 F.2d 1322, 1326 n.11, 206 USPQ 885, 889 n.11 (CCPA 1980); *In re Fouche*,

439 F.2d 1237, 1243, 169 USPQ 429, 434 (CCPA 1971)), Appellants submit that as claims 1-4, 11 and 12 have been shown to have "a specific, substantial, and credible utility", as detailed in **Section VIII(A)** above, the present rejection of claims 1-4, 11 and 12 under 35 U.S.C. § 112, first paragraph, cannot stand.

Appellants therefore submit that the rejection of claims 1-4, 11 and 12 under 35 U.S.C. § 112, first paragraph, must be overruled.

C. Is Claim 11 indefinite?

Appellants would like to first apologize in advance to the Board for having to brief this issue, which would have been obviated had Appellants previously submitted amendment been entered into this case, however as this Brief is in four month extension without an Advisory Action and Appelants have been unable to confirm that prior submitted amendments have been entered in this case, rather than incur additional fees and potential loss of patent term, Appellants believe it far more prudent to have elected to respond at this time.

Claim 11 stands rejected under 35 U.S.C. § 112, second paragraph, as being allegedly indefinite for recitation of the allegedly indefinite recitation of the phrase "comprising [a] nucleic acid sequence of Claim 4." The Examiner interprets this to connote some part of the sequence of the nucleic acid sequence of Claim 4 and that replacement of "a" with "the" will resolve the issue.

Appellants submit that all that is required under 35 U.S.C. § 112, second paragraph, is that the skilled artisan be apprised both of the utilization and scope of the invention (*Shatterproof Glass Corp. v. Libbey Owens Ford Co.*, 225 USPQ 634 (Fed. Cir. 1985)). The Federal Circuit has clearly stated in *S3 v. Nvidia* (259 F.3d 1364 (Fed. Cir. 2001)):

> The requirement that the claims "particularly point out and distinctly claim" the invention is met when a person experienced in the field of the invention would understand the scope of the subject matter that is patented when the claim is read in conjunction with the rest of the specification. "If the claims read in light of the specification reasonably appraise those skilled in the art of the scope of the invention, § 112 demands no more" (*Miles Labs., Inc. v. Shandon, Inc.*, 27 USPQ2d 1123, 1126 (Fed. Cir. 1993); *see also Union Pacific*

Resources Co. v. Chesapeake Energy Corp., 236 F.3d 684, 692, 57 USPQ2d 1293, 1297 (Fed. Cir. 2001); *North American Vaccine, Inc. v. American Cyanamid Co.*, F.3d 1571, 1579, 28 USPQ2d 1333, 1339 (Fed. Cir. 1993); *Hybritech, Inc. v. Monoclonal Antibodies*, 802 F.2d 1367, 1385, 231 USPQ 81, 94-95 (Fed. Cir. 1986).

Appellants understand that while traditional dependant claim construction would suggest that the use of the term "the" in Claim 11 would be technically correct. The present situation differs in that due to the degeneracy of the nucleic acid triplet codons, more than one nucleic acid sequence can encode a single amino acid sequence. This degeneracy has been known to those of skill in the art since the 1960s and is accepted by those of skill in the art. Thus while finite the a nucleotide sequence that encodes the amino acid sequence shown in SEQ ID NO 4", does not represent a single nucleic acid sequence. Therefore, the use of the word "the" in this context would have been grammatically incorrect.

While in no way agreeing with the Examiner's rejection, in order to advance this application more quickly towards allowance, Appellants submitted an amendment to Claim 11 in the belief that in all but the most tortured reading of the claim, one of skill in the art would readily understand that whether phrased as "the" nucleic acid or "a" nucleic acid were used in the context of Claim 11, it would be interpreted as meaning any of the nucleic acid sequences defined by Claim 4, that is to say any of the nucleic acid sequences that encodes the amino acid sequence shown in SEQ ID NO 4. Thus, those of skill in the art would have recognized both the utilization and scope of the invention as set forth in Claim 11, and therefore, Claim 11 meets the requirements of 35 U.S.C. § 112, second paragraph and this rejection should be overturned.

IX. APPENDIX

The claims involved in this appeal are as follows:

1. An isolated nucleic acid molecule comprising the nucleotide sequence of SEQ ID NO: 1.

2. An isolated nucleic acid molecule comprising a nucleotide sequence that:
   (a) encodes the amino acid sequence shown in SEQ ID NO: 2; and
   (b) hybridizes under stringent conditions to the nucleotide sequence of SEQ ID NO: 1 or the complement thereof.

3. An isolated nucleic acid molecule comprising a nucleotide sequence that encodes the amino acid sequence shown in SEQ ID NO: 2.

4. An isolated nucleic acid molecule comprising a nucleotide sequence that encodes the amino acid sequence shown in SEQ ID NO:4.

11. An expression vector comprising a nucleic acid sequence of Claim 4.

12. A cell comprising the expression vector of Claim 11.

X. CONCLUSION

Appellants respectfully submit that, in light of the foregoing arguments, the Final Action's conclusion that claims 1-4, 11 and 12 lack a patentable utility and are unusable by the skilled artisan due to a lack of patentable utility is unwarranted. It is therefore requested that the Board overturn the Final Action's rejections.

Respectfully submitted,

March 26, 2004
Date

Lance K. Ishimoto Reg. No. 41,866
Agent For Appellants

LEXICON GENETICS INCORPORATED
(281) 863-3399

Customer # 24231

```
 FASTA searches a protein or DNA sequence data bank
 version 3.3t05 March 30, 2000
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

/tmp/fastaCAANHaihS: 874 aa
>>SEQ ID NO:3 human semaphorin
 vs  /tmp/fastaDAAOHaihS library
searching /tmp/fastaDAAOHaihS library

    782 residues in     1 sequences

FASTA (3.34 January 2000) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 38, opt: 26, gap-pen: -12/ -2, width:  16
 Scan time:  0.017
The best scores are:                                          opt
gi|8978202|dbj|BAA98132.1| semaphorin sem2 [Homo    ( 782) 5450

>>gi|8978202|dbj|BAA98132.1| semaphorin sem2 [Homo sapie  (782 aa)
 initn: 5448 init1: 4266 opt: 5450
Smith-Waterman score: 5450;  99.872% identity in 782 aa overlap (94-874:1-782)

           70        80        90       100       110       120
SEQ    GGSRANYNRRPAGPEGGSAGRRQRCPQFPSMAPSAWAICWLLGGLLLHGGSSGPSPGPSV
                                       ::::::::::::::::::::::::::::
gi|897                                 MAPSAWAICWLLGGLLLHGGSSGPSPGPSV
                                               10        20        30

           130       140       150       160       170       180
SEQ    PRLRLSYRDLLSANRSAIFLGPQGSLNLQAMYLDEYRDRLFLGGLDALYSLRLDQAWPDP
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 PRLRLSYRDLLSANRSAIFLGPQGSLNLQAMYLDEYRDRLFLGGLDALYSLRLDQAWPDP
               40        50        60        70        80        90

           190       200       210       220       230       240
SEQ    REVLWPPQPGQREECVRKGRDPLTECANFVRVLQPHNRTHLLACGTGAFQPTCALITVGH
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 REVLWPPQPGQREECVRKGRDPLTECANFVRVLQPHNRTHLLACGTGAFQPTCALITVGH
              100       110       120       130       140       150

           250       260       270       280       290       300
SEQ    RGEHVLHLEPGSVESGRGRCPHEPSRPFASTFIDGELYTGLTADFLGREAMIFRSGGPRP
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 RGEHVLHLEPGSVESGRGRCPHEPSRPFASTFIDGELYTGLTADFLGREAMIFRSGGPRP
              160       170       180       190       200       210

           310       320       330       340       350       360
SEQ    ALRSDSDQSLLHDPRFVMAARIPENSDQDNDKVYFFFSETVPSPDGGSNHVTVSRVGRVC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 ALRSDSDQSLLHDPRFVMAARIPENSDQDNDKVYFFFSETVPSPDGGSNHVTVSRVGRVC
              220       230       240       250       260       270

           370       380       390       400       410       420
SEQ    VNDAGGQRVLVNKWSTFLKARLVCSVPGPGGAETHFDQLEDVFLLWPKAGKSLEVYALFS
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 VNDAGGQRVLVNKWSTFLKARLVCSVPGPGGAETHFDQLEDVFLLWPKAGKSLEVYALFS
              280       290       300       310       320       330

           430       440       450       460       470       480
SEQ    TVSAVFQGFAVCVYHMADIWEVFNGPFAHRDGPQHQWGPYGGKVPFPRPGVCPSKMTAQP
```

```
          :::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 TVSAVFQGFAVCVYHMADIWEVFNGPFAHRDGPQHQWGPYGGKVPFPRPGVCPSKMTAQP
               340       350       360       370       380       390

           490       500       510       520       530       540
SEQ    GRPFGSTKDYPDEVLQFARAHPLMFWPVRPRHGRPVLVKTHLAQQLHQIVVDRVEAEDGT
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GRPFGSTKDYPDEVLQFARAHPLMFWPVRPRHGRPVLVKTHLAQQLHQIVVDRVEAEDGT
               400       410       420       430       440       450

         550       560       570       580       590       600
SEQ    YDVIFLGTDSGSVLKVIALQAGGSAEPEEVVLEELQVFKVPTPITEMEISVKRQMLYVGS
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 YDVIFLGTDSGSVLKVIALQAGGSAEPEEVVLEELQVFKVPTPITEMEISVKRQMLYVGS
               460       470       480       490       500       510

           610       620       630       640       650       660
SEQ    RLGVAQLRLHQCETYGTACAECCLARDPYCAWDGASCTHYRPSLGKRRFRRQDIRHGNPA
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 RLGVAQLRLHQCETYGTACAECCLARDPYCAWDGASCTHYRPSLGKRRFRRQDIRHGNPA
               520       530       540       550       560       570

           670       680       690       700       710       720
SEQ    LQCLGQSQEEEAVGLVAATMVYGTEHNSTFLECLPKSP-AAVRWLLQRPGDEGPDQVKTD
       :::::::::::::::::::::::::::::::::::: :::::::::::::::::::::
gi|897 LQCLGQSQEEEAVGLVAATMVYGTEHNSTFLECLPKSPQAAVRWLLQRPGDEGPDQVKTD
               580       590       600       610       620       630

           730       740       750       760       770       780
SEQ    ERVLHTERGLLFRRLSRFDAGTYTCTTLEHGFSQTVVRLALVVIVASQLDNLFPPEPKPE
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 ERVLHTERGLLFRRLSRFDAGTYTCTTLEHGFSQTVVRLALVVIVASQLDNLFPPEPKPE
               640       650       660       670       680       690

           790       800       810       820       830       840
SEQ    EPPARGGLASTPPKAWYKDILQLIGFANLPRVDEYCERVWCRGTTECSGCFRSRSRGKQA
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 EPPARGGLASTPPKAWYKDILQLIGFANLPRVDEYCERVWCRGTTECSGCFRSRSRGKQA
               700       710       720       730       740       750

           850       860       870
SEQ    RGKSWAGLELGKKMKSRVHAEHNRTPREVEAT
       ::::::::::::::::::::::::::::::::
gi|897 RGKSWAGLELGKKMKSRVHAEHNRTPREVEAT
               760       770       780


874 residues in 1 query  sequences
782 residues in 1 library sequences
 Scomplib [version 3.3t05 March 30, 2000]
 start: Mon Feb 10 16:12:17 2003 done: Mon Feb 10 16:12:17 2003
 Scan time:  0.017 Display time:  0.933

Function used was FASTA
```

```
 FASTA searches a protein or DNA sequence data bank
 version 3.3t05 March 30, 2000
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

/tmp/fastaCAAWTaiDv: 2349 nt
 >LEX151 SEQ ID NO:3
 vs  /tmp/fastaDAAXTaiDv library
searching /tmp/fastaDAAXTaiDv library

   4700 residues in     1 sequences

FASTA (3.34 January 2000) function [optimized, +5/-4 matrix (5:-4)] ktup: 6
 join: 74, opt: 59, gap-pen: -16/ -4, width:  16
 Scan time:  0.100
The best scores are:                                                  opt
gi|8978201|dbj|AB029496.1| Homo sapiens mRNA f (4700) [f] 11742
gi|8978201|dbj|AB029496.1| Homo sapiens mRNA f (4700) [r]   78

>>gi|8978201|dbj|AB029496.1| Homo sapiens mRNA for semap  (4700 nt)
 initn: 11742 init1: 11742 opt: 11742
 99.957% identity in 2349 nt overlap (1-2349:1-2349)

               10        20        30        40        50        60
LEX151 ATGGCCCCCTCGGCCTGGGCCATTTGCTGGCTGCTAGGGGGCCTCCTGCTCCATGGGGGT
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 ATGGCCCCCTCGGCCTGGGCCATTTGCTGGCTGCTAGGGGGCCTCCTGCTCCATGGGGGT
               10        20        30        40        50        60

               70        80        90       100       110       120
LEX151 AGCTCTGGCCCCAGCCCCGGCCCCAGTGTGCCCCGCCTGCGGCTCTCCTACCGAGACCTC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 AGCTCTGGCCCCAGCCCCGGCCCCAGTGTGCCCCGCCTGCGGCTCTCCTACCGAGACCTC
               70        80        90       100       110       120

              130       140       150       160       170       180
LEX151 CTGTCTGCCAACCGCTCTGCCATCTTTCTGGGCCCCCAGGGCTCCCTGAACCTCCAGGCC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 CTGTCTGCCAACCGCTCTGCCATCTTTCTGGGCCCCCAGGGCTCCCTGAACCTCCAGGCC
              130       140       150       160       170       180

              190       200       210       220       230       240
LEX151 ATGTACCTAGATGAGTACCGAGACCGCCTCTTTCTGGGTGGCCTGGACGCCCTCTACTCT
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 ATGTACCTAGATGAGTACCGAGACCGCCTCTTTCTGGGTGGCCTGGACGCCCTCTACTCT
              190       200       210       220       230       240

              250       260       270       280       290       300
LEX151 CTGCGGCTGGACCAGGCATGGCCAGATCCCCGGGAGGTCCTGTGGCCACCGCAGCCAGGA
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 CTGCGGCTGGACCAGGCATGGCCAGATCCCCGGGAGGTCCTGTGGCCACCGCAGCCAGGA
              250       260       270       280       290       300

              310       320       330       340       350       360
LEX151 CAGAGGGAGGAGTGTGTTCGAAAGGGAAGAGATCCTTTGACAGAGTGCGCCAACTTCGTG
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 CAGAGGGAGGAGTGTGTTCGAAAGGGAAGAGATCCTTTGACAGAGTGCGCCAACTTCGTG
              310       320       330       340       350       360

              370       380       390       400       410       420
```

Compare Genomic Sequences

```
LEX151 CGGGTGCTACAGCCTCACAACCGGACCCACCTGCTAGCCTGTGGCACTGGGGCCTTCCAG
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 CGGGTGCTACAGCCTCACAACCGGACCCACCTGCTAGCCTGTGGCACTGGGGCCTTCCAG
              370       380       390       400       410       420

              430       440       450       460       470       480
LEX151 CCCACCTGTGCCCTCATCACAGTTGGCCACCGTGGGGAGCATGTGCTCCACCTGGAGCCT
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 CCCACCTGTGCCCTCATCACAGTTGGCCACCGTGGGGAGCATGTGCTCCACCTGGAGCCT
              430       440       450       460       470       480

              490       500       510       520       530       540
LEX151 GGCAGTGTGGAAAGTGGCCGGGGGCGGTGCCCTCACGAGCCCAGCCGTCCCTTTGCCAGC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GGCAGTGTGGAAAGTGGCCGGGGGCGGTGCCCTCACGAGCCCAGCCGTCCCTTTGCCAGC
              490       500       510       520       530       540

              550       560       570       580       590       600
LEX151 ACCTTCATAGACGGGGAGCTGTACACGGGTCTCACTGCTGACTTCCTGGGGCGAGAGGCC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 ACCTTCATAGACGGGGAGCTGTACACGGGTCTCACTGCTGACTTCCTGGGGCGAGAGGCC
              550       560       570       580       590       600

              610       620       630       640       650       660
LEX151 ATGATCTTCCGAAGTGGAGGTCCTCGGCCAGCTCTGCGTTCCGACTCTGACCAGAGTCTC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 ATGATCTTCCGAAGTGGAGGTCCTCGGCCAGCTCTGCGTTCCGACTCTGACCAGAGTCTC
              610       620       630       640       650       660

              670       680       690       700       710       720
LEX151 TTGCACGACCCCCGGTTTGTGATGGCCGCCCGGATCCCTGAGAACTCTGACCAGGACAAT
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 TTGCACGACCCCCGGTTTGTGATGGCCGCCCGGATCCCTGAGAACTCTGACCAGGACAAT
              670       680       690       700       710       720

              730       740       750       760       770       780
LEX151 GACAAGGTGTACTTCTTCTTCTCGGAGACGGTCCCCTCGCCCGATGGTGGCTCGAACCAT
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GACAAGGTGTACTTCTTCTTCTCGGAGACGGTCCCCTCGCCCGATGGTGGCTCGAACCAT
              730       740       750       760       770       780

              790       800       810       820       830       840
LEX151 GTCACTGTCAGCCGCGTGGGCCGCGTCTGCGTGAATGATGCTGGGGGCCAGCGGGTGCTG
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GTCACTGTCAGCCGCGTGGGCCGCGTCTGCGTGAATGATGCTGGGGGCCAGCGGGTGCTG
              790       800       810       820       830       840

              850       860       870       880       890       900
LEX151 GTGAACAAATGGAGCACTTTCCTCAAGGCCAGGCTGGTCTGCTCGGTGCCCGGCCCTGGT
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GTGAACAAATGGAGCACTTTCCTCAAGGCCAGGCTGGTCTGCTCGGTGCCCGGCCCTGGT
              850       860       870       880       890       900

              910       920       930       940       950       960
LEX151 GGTGCCGAGACCCACTTTGACCAGCTAGAGGATGTGTTCCTGCTGTGGCCCAAGGCCGGG
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GGTGCCGAGACCCACTTTGACCAGCTAGAGGATGTGTTCCTGCTGTGGCCCAAGGCCGGG
              910       920       930       940       950       960

              970       980       990      1000      1010      1020
```

Compare Genomic Sequences

```
LEX151 AAGAGCCTCGAGGTGTACGCGCTGTTCAGCACCGTCAGTGCCGTGTTCCAGGGCTTCGCC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 AAGAGCCTCGAGGTGTACGCGCTGTTCAGCACCGTCAGTGCCGTGTTCCAGGGCTTCGCC
            970       980       990      1000      1010      1020

            1030      1040      1050      1060      1070      1080
LEX151 GTCTGTGTGTACCACATGGCAGACATCTGGGAGGTTTTCAACGGGCCCTTTGCCCACCGA
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GTCTGTGTGTACCACATGGCAGACATCTGGGAGGTTTTCAACGGGCCCTTTGCCCACCGA
            1030      1040      1050      1060      1070      1080

            1090      1100      1110      1120      1130      1140
LEX151 GATGGGCCTCAGCACCAGTGGGGGCCCTATGGGGGCAAGGTGCCCTTCCCTCGCCCTGGC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GATGGGCCTCAGCACCAGTGGGGGCCCTATGGGGGCAAGGTGCCCTTCCCTCGCCCTGGC
            1090      1100      1110      1120      1130      1140

            1150      1160      1170      1180      1190      1200
LEX151 GTGTGCCCCAGCAAGATGACCGCACAGCCAGGACGGCCTTTTGGCAGCACCAAGGACTAC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GTGTGCCCCAGCAAGATGACCGCACAGCCAGGACGGCCTTTTGGCAGCACCAAGGACTAC
            1150      1160      1170      1180      1190      1200

            1210      1220      1230      1240      1250      1260
LEX151 CCAGATGAGGTGCTGCAGTTTGCCCGAGCCCACCCCCTCATGTTCTGGCCTGTGCGGCCT
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 CCAGATGAGGTGCTGCAGTTTGCCCGAGCCCACCCCCTCATGTTCTGGCCTGTGCGGCCT
            1210      1220      1230      1240      1250      1260

            1270      1280      1290      1300      1310      1320
LEX151 CGACATGGCCGCCCTGTCCTTGTCAAGACCCACCTGGCCCAGCAGCTACACCAGATCGTG
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 CGACATGGCCGCCCTGTCCTTGTCAAGACCCACCTGGCCCAGCAGCTACACCAGATCGTG
            1270      1280      1290      1300      1310      1320

            1330      1340      1350      1360      1370      1380
LEX151 GTGGACCGCGTGGAGGCAGAGGATGGGACCTACGATGTCATTTTCCTGGGGACTGACTCA
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GTGGACCGCGTGGAGGCAGAGGATGGGACCTACGATGTCATTTTCCTGGGGACTGACTCA
            1330      1340      1350      1360      1370      1380

            1390      1400      1410      1420      1430      1440
LEX151 GGGTCTGTGCTCAAAGTCATCGCTCTCCAGGCAGGGGGCTCAGCTGAACCTGAGGAAGTG
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GGGTCTGTGCTCAAAGTCATCGCTCTCCAGGCAGGGGGCTCAGCTGAACCTGAGGAAGTG
            1390      1400      1410      1420      1430      1440

            1450      1460      1470      1480      1490      1500
LEX151 GTTCTGGAGGAGCTCCAGGTGTTTAAGGTGCCAACACCTATCACCGAAATGGAGATCTCT
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GTTCTGGAGGAGCTCCAGGTGTTTAAGGTGCCAACACCTATCACCGAAATGGAGATCTCT
            1450      1460      1470      1480      1490      1500

            1510      1520      1530      1540      1550      1560
LEX151 GTCAAAAGGCAAATGCTATACGTGGGCTCTCGGCTGGGTGTGGCCCAGCTGCGGCTGCAC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GTCAAAAGGCAAATGCTATACGTGGGCTCTCGGCTGGGTGTGGCCCAGCTGCGGCTGCAC
            1510      1520      1530      1540      1550      1560

            1570      1580      1590      1600      1610      1620
```

```
LEX151 CAATGTGAGACTTACGGCACTGCCTGTGCAGAGTGCTGCCTGGCCCGGGACCCATACTGT
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 CAATGTGAGACTTACGGCACTGCCTGTGCAGAGTGCTGCCTGGCCCGGGACCCATACTGT
            1570      1580      1590      1600      1610      1620

            1630      1640      1650      1660      1670      1680
LEX151 GCCTGGGATGGTGCCTCCTGTACCCACTACCGCCCCAGCCTTGGCAAGCGCCGGTTCCGC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GCCTGGGATGGTGCCTCCTGTACCCACTACCGCCCCAGCCTTGGCAAGCGCCGGTTCCGC
            1630      1640      1650      1660      1670      1680

            1690      1700      1710      1720      1730      1740
LEX151 CGGCAGGACATCCGGCACGGCAACCCTGCCCTGCAGTGCCTGGGCCAGAGCCAGGAAGAA
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 CGGCAGGACATCCGGCACGGCAACCCTGCCCTGCAGTGCCTGGGCCAGAGCCAGGAAGAA
            1690      1700      1710      1720      1730      1740

            1750      1760      1770      1780      1790      1800
LEX151 GAGGCAGTGGGACTTGTGGCAGCCACCATGGTCTACGGCACGGAGCACAATAGCACCTTC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GAGGCAGTGGGACTTGTGGCAGCCACCATGGTCTACGGCACGGAGCACAATAGCACCTTC
            1750      1760      1770      1780      1790      1800

            1810      1820      1830      1840      1850      1860
LEX151 CTGGAGTGCCTGCCCAAGTCTCCCCARGCTGCTGTGCGCTGGCTCTTGCAGAGGCCAGGG
       ::::::::::::::::::::::::.::::::::::::::::::::::::::::::::::::
gi|897 CTGGAGTGCCTGCCCAAGTCTCCCCAGGCTGCTGTGCGCTGGCTCTTGCAGAGGCCAGGG
            1810      1820      1830      1840      1850      1860

            1870      1880      1890      1900      1910      1920
LEX151 GATGAGGGGCCTGACCAGGTGAAGACGGACGAGCGAGTCTTGCACACGGAGCGGGGGCTG
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GATGAGGGGCCTGACCAGGTGAAGACGGACGAGCGAGTCTTGCACACGGAGCGGGGGCTG
            1870      1880      1890      1900      1910      1920

            1930      1940      1950      1960      1970      1980
LEX151 CTGTTCCGCAGGCTTAGCCGTTTCGATGCGGGCACCTACACCTGCACCACTCTGGAGCAT
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 CTGTTCCGCAGGCTTAGCCGTTTCGATGCGGGCACCTACACCTGCACCACTCTGGAGCAT
            1930      1940      1950      1960      1970      1980

            1990      2000      2010      2020      2030      2040
LEX151 GGCTTCTCCCAGACTGTGGTCCGCCTGGCTCTGGTGGTGATTGTGGCCTCACAGCTGGAC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GGCTTCTCCCAGACTGTGGTCCGCCTGGCTCTGGTGGTGATTGTGGCCTCACAGCTGGAC
            1990      2000      2010      2020      2030      2040

            2050      2060      2070      2080      2090      2100
LEX151 AACCTGTTCCCTCCGGAGCCAAAGCCAGAGGAGCCCCCAGCCCGGGGAGGCCTGGCTTCC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 AACCTGTTCCCTCCGGAGCCAAAGCCAGAGGAGCCCCCAGCCCGGGGAGGCCTGGCTTCC
            2050      2060      2070      2080      2090      2100

            2110      2120      2130      2140      2150      2160
LEX151 ACCCCACCCAAGGCCTGGTACAAGGACATCCTGCAGCTCATTGGCTTCGCCAACCTGCCC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 ACCCCACCCAAGGCCTGGTACAAGGACATCCTGCAGCTCATTGGCTTCGCCAACCTGCCC
            2110      2120      2130      2140      2150      2160

            2170      2180      2190      2200      2210      2220
```

```
LEX151 CGGGTGGATGAGTACTGTGAGCGCGTGTGGTGCAGGGGCACCACGGAATGCTCAGGCTGC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 CGGGTGGATGAGTACTGTGAGCGCGTGTGGTGCAGGGGCACCACGGAATGCTCAGGCTGC
             2170      2180      2190      2200      2210      2220

             2230      2240      2250      2260      2270      2280
LEX151 TTCCGGAGCCGGAGCCGGGGCAAGCAGGCCAGGGGCAAGAGCTGGGCAGGGCTGGAGCTA
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 TTCCGGAGCCGGAGCCGGGGCAAGCAGGCCAGGGGCAAGAGCTGGGCAGGGCTGGAGCTA
             2230      2240      2250      2260      2270      2280

             2290      2300      2310      2320      2330      2340
LEX151 GGCAAGAAGATGAAGAGCCGGGTGCATGCCGAGCACAATCGGACGCCCCGGGAGGTGGAG
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GGCAAGAAGATGAAGAGCCGGGTGCATGCCGAGCACAATCGGACGCCCCGGGAGGTGGAG
             2290      2300      2310      2320      2330      2340


LEX151 GCCACGTAG
       :::::::::
gi|897 GCCACGTAGAAGGGGGCAGAGGAGGGGTGGTCAGGATGGGCTGGGGGGCCCACTAGCAGC
             2350      2360      2370      2380      2390      2400

>>gi|8978201|dbj|AB029496.1| Homo sapiens mRNA for semap  (4700 nt)
rev-comp initn: 136 init1:  78 opt:  78
 85.714% identity in 21 nt overlap (875-855:476-496)

           900       890       880       870       860       850
LEX15- GCACCACCAGGGCCGGGCACCGAGCAGACCAGCCTGGCCTTGAGGAAAGTGCTCCATTTG
                                        :::::::: :: ::::::::
gi|897 GCCACCGTGGGGAGCATGTGCTCCACCTGGAGCCTGGCAGTGTGGAAAGTGGCCGGGGGC
           450       460       470       480       490       500

           840       830       820       810       800       790
LEX15- TTCACCAGCACCCGCTGGCCCCCAGCATCATTCACGCAGACGCGGCCCACGCGGCTGACA

gi|897 GGTGCCCTCACGAGCCCAGCCGTCCCTTTGCCAGCACCTTCATAGACGGGGAGCTGTACA
           510       520       530       540       550       560


2349 residues in 1 query   sequences
4700 residues in 1 library sequences
 Scomplib [version 3.3t05 March 30, 2000]
 start: Fri Sep 19 13:51:42 2003 done: Fri Sep 19 13:51:42 2003
 Scan time:  0.100 Display time:  0.150

Function used was FASTA
```

☐ 1: AB029496. Homo sapiens mRNA...[gi:8978201] Links

```
LOCUS       AB029496                4700 bp    mRNA    linear   PRI 07-JUL-2000
DEFINITION  Homo sapiens mRNA for semaphorin sem2, complete cds.
ACCESSION   AB029496
VERSION     AB029496.1  GI:8978201
KEYWORDS    semaphorin sem2.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 4700)
  AUTHORS   Seki,N., Hattori,A., Hayashi,A., Kozuma,S., Muramatsu,M.,
            Miyajima,N. and Saito,T.
  TITLE     Human semaphorin
  JOURNAL   Published Only in DataBase (2000)
REFERENCE   2  (bases 1 to 4700)
  AUTHORS   Seki,N., Hattori,A., Hayashi,A., Kozuma,S., Muramatsu,M.,
            Miyajima,N. and Saito,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (01-JUL-1999) Toshiyuki Saito, National Institute of
            Radiological Sciences, Genome Research Group; Inage-ku Anagawa
            4-9-1, Chiba, Chiba 263-8555, Japan (E-mail:t_saito@nirs.go.jp,
            Tel:81-43-201-3135, Fax:81-43-251-9818)
FEATURES             Location/Qualifiers
     source          1..4700
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
     gene            1..4700
                     /gene="sem2"
     CDS             1..2349
                     /gene="sem2"
                     /codon_start=1
                     /product="semaphorin sem2"
                     /protein_id="BAA98132.1"
                     /db_xref="GI:8978202"
                     /translation="MAPSAWAICWLLGGLLLHGGSSGPSPGPSVPRLRLSYRDLLSAN
                     RSAIFLGPQGSLNLQAMYLDEYRDRLFLGGLDALYSLRLDQAWPDPREVLWPPQPGQR
                     EECVRKGRDPLTECANFVRVLQPHNRTHLLACGTGAFQPTCALITVGHRGEHVLHLEP
                     GSVESGRGRCPHEPSRPFASTFIDGELYTGLTADFLGREAMIFRSGGPRPALRSDSDQ
                     SLLHDPRFVMAARIPENSDQDNDKVYFFFSETVPSPDGGSNHVTVSRVGRVCVNDAGG
                     QRVLVNKWSTFLKARLVCSVPGPGGAETHFDQLEDVFLLWPKAGKSLEVYALFSTVSA
                     VFQGFAVCVYHMADIWEVFNGPFAHRDGPQHQWGPYGGKVPFPRPGVCPSKMTAQPGR
                     PFGSTKDYPDEVLQFARAHPLMFWPVRPRHGRPVLVKTHLAQQLHQIVVDRVEAEDGT
                     YDVIFLGTDSGSVLKVIALQAGGSAEPEEVVLEELQVFKVPTPITEMEISVKRQMLYV
                     GSRLGVAQLRLHQCETYGTACAECCLARDPYCAWDGASCTHYRPSLGKRRFRRQDIRH
                     GNPALQCLGQSQEEEAVGLVAATMVYGTEHNSTFLECLPKSPQAAVRWLLQRPGDEGP
                     DQVKTDERVLHTERGLLFRRLSRFDAGTYTCTTLEHGFSQTVVRLALVVIVASQLDNL
```

FPPEPKPEEPPARGGLASTPPKAWYKDILQLIGFANLPRVDEYCERVWCRGTTECSGC
FRSRSRGKQARGKSWAGLELGKKMKSRVHAEHNRTPREVEAT"

BASE COUNT 972 a 1307 c 1467 g 954 t

ORIGIN

```
      1 atggcccccc cggcctgggc catttgctgg ctgctagggg gcctcctgct ccatgggggt
     61 agctctggcc ccagccccgg ccccagtgtg ccccgcctgc ggctctccta ccgagacctc
    121 ctgtctgcca accgctctgc catctttctg ggcccccagg gctccctgaa cctccaggcc
    181 atgtacctag atgagtaccg agaccgcctc tttctgggtg gcctggacgc cctctactct
    241 ctgcggctgg accaggcatg gccagatccc cgggaggtcc tgtggccacc gcagccagga
    301 cagagggagg agtgtgttcg aaagggaaga gatcctttga cagagtgcgc caacttcgtg
    361 cgggtgctac agcctcacaa ccggacccac ctgctagcct gtggcactgg ggccttccag
    421 cccacctgtg ccctcatcac agttggccac cgtggggagc atgtgctcca cctggagcct
    481 ggcagtgtgg aaagtggccg ggggcggtgc cctcacgagc ccagccgtcc ctttgccagc
    541 accttcatag acggggagct gtacacgggt ctcactgctg acttcctggg gcgagaggcc
    601 atgatcttcc gaagtggagg tcctcggcca gctctgcgtt ccgactctga ccagagtctc
    661 ttgcacgacc cccggtttgt gatggccgcc cggatccctg agaactctga ccaggacaat
    721 gacaaggtgt acttcttctt ctcggagacg gtcccctcgc ccgatggtgg ctcgaaccat
    781 gtcactgtca gccgcgtggg ccgcgtctgc gtgaatgatg ctgggggcca gcgggtgctg
    841 gtgaacaaat ggagcacttt cctcaaggcc aggctggtct gctcggtgcc cggccctggt
    901 ggtgccgaga cccactttga ccagctagag gatgtgttcc tgctgtggcc caaggccggg
    961 aagagcctcg aggtgtacgc gctgttcagc accgtcagtg ccgtgttcca gggcttcgcc
   1021 gtctgtgtgt accacatggc agacatctgg gaggttttca acgggccctt tgcccaccga
   1081 gatgggcctc agcaccagtg ggggccctat gggggcaagg tgcccttccc tcgccctggc
   1141 gtgtgcccca gcaagatgac cgcacagcca ggacggcctt ttggcagcac caaggactac
   1201 ccagatgagg tgctgcagtt tgcccgagcc cacccctca tgttctggcc tgtgcggcct
   1261 cgacatggcc gccctgtcct tgtcaagacc cacctggccc agcagctaca ccagatcgtg
   1321 gtggaccgcg tggaggcaga ggatgggacc tacgatgtca ttttcctggg gactgactca
   1381 gggtctgtgc tcaaagtcat cgctctccag gcagggggct cagctgaacc tgaggaagtg
   1441 gttctggagg agctccaggt gtttaaggtg ccaacaccta tcaccgaaat ggagatctct
   1501 gtcaaaaggc aaatgctata cgtgggctct cggctgggtg tggcccagct gcggctgcac
   1561 caatgtgaga cttacggcac tgcctgtgca gagtgctgcc tggcccggga cccatactgt
   1621 gcctgggatg gtgcctcctg tacccactac cgccccagcc ttggcaagcg ccggttccgc
   1681 cggcaggaca tccggcacgg caaccctgcc ctgcagtgcc tgggccagag ccaggaagaa
   1741 gaggcagtgg gacttgtggc agccaccatg gtctacggca cggagcacaa tagcaccttc
   1801 ctggagtgcc tgcccaagtc tccccaggct gctgtgcgct ggctcttgca gaggccaggg
   1861 gatgaggggc ctgaccaggt gaagacggac gagcgagtct tgcacacgga gcgggggctg
   1921 ctgttccgca ggcttagccg tttcgatgcg ggcacctaca cctgcaccac tctggagcat
   1981 ggcttctccc agactgtggt ccgcctggct ctggtggtga ttgtggcctc acagctggac
   2041 aacctgttcc ctccggagcc aaagccagag gagcccccag cccggggagg cctggcttcc
   2101 accccaccca aggcctggta caaggacatc ctgcagctca ttggcttcgc caacctgccc
   2161 cgggtggatg agtactgtga gcgcgtgtgg tgcaggggca ccacggaatg ctcaggctgc
   2221 ttccggagcc ggagccgggg caagcaggcc aggggcaaga gctgggcagg gctggagcta
   2281 ggcaagaaga tgaagagccg ggtgcatgcc gagcacaatc ggacgccccg ggaggtggag
   2341 gccacgtaga agggggcaga ggaggggtgg tcaggatggg ctgggggccc cactagcagc
   2401 ccccagcatc tcccacccac ccagctaggg cagaggggtc aggatgtctg tttgcctctt
   2461 agagacaggt gtctctgccc ccacaccgct actggggtct aatggagggg ctgggttctt
   2521 gaagcctgtt ccctgccctt ctctgtgctc ttagacccag ctggagccag caccctctgg
   2581 ctgctggcag ccccaaggga tctgccattt gttctcagag atggcctggc ttccgcaaca
   2641 catttccggg tgtgcccaga ggcaagaggg ttgggtggtt ctttcccagc ctacagaaca
   2701 atggccattc tgagtgaccc tcagagtggg tgtgtgggtg cgtctagggg gtatcccggt
   2761 agggggcctg cagggagcca gagggtggaa atggcctcta agctagcacc ccgtaagaag
   2821 agcctacctg accgacttgg ggagggaaca cagaggtgtt gggaaggtgg agcaacaatg
   2881 cacctccccc cctgtcgcgc cgtgatatct tggtggctcc ctgccactgc ccaccgcctc
   2941 ttctccatct gagaatcacg gagaggtgta gataatctag aggcatagac tgctagagcc
   3001 cccagggatc tggggtggtc agggctcagg cttcactttg taaaccaggt gggggcatct
   3061 cacagcctga cttcccttcc ccaggccagg gttgctggga tgcctgcccc tcctgagagg
   3121 acccccctccc cattgtcagg ctctccatgt ccacgagcgg ggaggggtgg gttctggggc
   3181 attgttgtcc cttgtgtctg tggactagag atagggtggg ggagctgggg aagggtgcag
   3241 gcgggaagag tgggctgtct ttcccagggt gatgcaagca tgccgcagcc ctggaggctg
   3301 ggaatgtgga ggctctgtga gccctgcagc cctcagaatc agggccaggg atgcagaaga
```

```
3361 ttgagaggat atggagatgg atagagggca ggagaccctt aggatagatt gtgggaccca
3421 ggcaggaaca ggtgtccaca agaactcagg atggcatcag ttagctcaga agccacctgg
3481 aagacccagt gtttccatct ctggaatctc tgttttatgc taaatggatt taggaagact
3541 gtttttcttt taaggggaa acaaggtaga gaaaaggacg aagaagtgta agtcccgctg
3601 attctcgggg gtaaggctcg gatggcaagg acgcgttctg cctgggcatg taggggaggt
3661 gtttttgcca tcaccagttt ctcaggctgg ggagcacaga ggggaggagg aggactaaat
3721 gaaaagttgt tcccagcctg cacatgaaca cattcatgac acacaaaact ggctggaagg
3781 agataagagc actgggtttg agattccctc cattaaaaca accaagacaa agaaaggagg
3841 ggaaaaaaag ataaaaagca agccagggtt ccctgcccta ttgaaactca aacccagact
3901 gccttgggtt ttatctttcc cttacccctg gcacctccag agaactggga cctgaaatag
3961 tccctccgtt ctccccttg accatgtaat aaatgaacca gaagcactga gattaaccta
4021 tcaacgccct gagaagcctt ccagcctgcg gtgctgtctg ctgggaggtc agctggtcaa
4081 ggcagaggag gagaggagga aaggatgggg gctgaagagc agaagggagg ggagacagag
4141 gggattaaag aggggaggag agagtgcaga gctccaggaa agggtatcag agctgcagcc
4201 agctctgccc tctaccctag ggaggccaga aagacacaaa cagccctccg ggcctttacg
4261 ctggactctg gcttggcagg ctccaggcag ggtcctctgg gaagttactc tagaaaacga
4321 agggaggagg agcacaagat cctcagcaac gaacacctgc acttagaaaa agtggacagc
4381 ttctgccaac cacaccctac ccatggtact gtatgctatt aactcctgga aacgccccgt
4441 aaatgcgagt tgtttttgta tttgtgtgtt gagatgggcc ttgtggtttc tctgtactca
4501 gagcacattt cttgtaatta ctattgttat ttttattgtc atgactgccc ctgagctctg
4561 gtgagaaaag ctgaatttac aaggaaaggg atgaagttaa tatttgcatc acataattat
4621 atcattactg tgtatctgtg tattgtacta aatggactga tgctgcgcac atgagctgaa
4681 aatgaagagc cctcccatcc
//
```

```
 FASTA searches a protein or DNA sequence data bank
 version 3.3t05 March 30, 2000
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

/tmp/fastaGAAL1aqDv: 2628 nt
 >LEX151 SEQ ID NO:1
 vs  /tmp/fastaHAAM1aqDv library
searching /tmp/fastaHAAM1aqDv library

   4700 residues in     1 sequences

FASTA (3.34 January 2000) function [optimized, +5/-4 matrix (5:-4)] ktup: 6
 join: 77, opt: 62, gap-pen: -16/ -4, width:  16
 Scan time:  0.117
The best scores are:                                                opt
gi|8978201|dbj|AB029496.1| Homo sapiens mRNA f (4700) [f] 11742
gi|8978201|dbj|AB029496.1| Homo sapiens mRNA f (4700) [r]   95

>>gi|8978201|dbj|AB029496.1| Homo sapiens mRNA for semap  (4700 nt)
 initn: 11742 init1: 11742 opt: 11742
 99.957% identity in 2349 nt overlap (280-2628:1-2349)

     250       260       270       280       290       300
LEX151 AGGCGGCAGCGGTGCCCTCAGTTCCCCAGCATGGCCCCCTCGGCCTGGGCCATTTGCTGG
                                     ::::::::::::::::::::::::::::
gi|897                               ATGGCCCCCTCGGCCTGGGCCATTTGCTGG
                                             10        20        30

     310       320       330       340       350       360
LEX151 CTGCTAGGGGGCCTCCTGCTCCATGGGGGTAGCTCTGGCCCCAGCCCCGGCCCCAGTGTG
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 CTGCTAGGGGGCCTCCTGCTCCATGGGGGTAGCTCTGGCCCCAGCCCCGGCCCCAGTGTG
               40        50        60        70        80        90

     370       380       390       400       410       420
LEX151 CCCCGCCTGCGGCTCTCCTACCGAGACCTCCTGTCTGCCAACCGCTCTGCCATCTTTCTG
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 CCCCGCCTGCGGCTCTCCTACCGAGACCTCCTGTCTGCCAACCGCTCTGCCATCTTTCTG
              100       110       120       130       140       150

     430       440       450       460       470       480
LEX151 GGCCCCCAGGGCTCCCTGAACCTCCAGGCCATGTACCTAGATGAGTACCGAGACCGCCTC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GGCCCCCAGGGCTCCCTGAACCTCCAGGCCATGTACCTAGATGAGTACCGAGACCGCCTC
              160       170       180       190       200       210

     490       500       510       520       530       540
LEX151 TTTCTGGGTGGCCTGGACGCCCTCTACTCTCTGCGGCTGGACCAGGCATGGCCAGATCCC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 TTTCTGGGTGGCCTGGACGCCCTCTACTCTCTGCGGCTGGACCAGGCATGGCCAGATCCC
              220       230       240       250       260       270

     550       560       570       580       590       600
LEX151 CGGGAGGTCCTGTGGCCACCGCAGCCAGGACAGAGGGAGGAGTGTGTTCGAAAGGGAAGA
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 CGGGAGGTCCTGTGGCCACCGCAGCCAGGACAGAGGGAGGAGTGTGTTCGAAAGGGAAGA
              280       290       300       310       320       330

     610       620       630       640       650       660
```

```
LEX151 GATCCTTTGACAGAGTGCGCCAACTTCGTGCGGGTGCTACAGCCTCACAACCGGACCCAC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GATCCTTTGACAGAGTGCGCCAACTTCGTGCGGGTGCTACAGCCTCACAACCGGACCCAC
             340       350       360       370       380       390

    670       680       690       700       710       720
LEX151 CTGCTAGCCTGTGGCACTGGGGCCTTCCAGCCCACCTGTGCCCTCATCACAGTTGGCCAC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 CTGCTAGCCTGTGGCACTGGGGCCTTCCAGCCCACCTGTGCCCTCATCACAGTTGGCCAC
             400       410       420       430       440       450

    730       740       750       760       770       780
LEX151 CGTGGGGAGCATGTGCTCCACCTGGAGCCTGGCAGTGTGGAAAGTGGCCGGGGGCGGTGC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 CGTGGGGAGCATGTGCTCCACCTGGAGCCTGGCAGTGTGGAAAGTGGCCGGGGGCGGTGC
             460       470       480       490       500       510

    790       800       810       820       830       840
LEX151 CCTCACGAGCCCAGCCGTCCCTTTGCCAGCACCTTCATAGACGGGGAGCTGTACACGGGT
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 CCTCACGAGCCCAGCCGTCCCTTTGCCAGCACCTTCATAGACGGGGAGCTGTACACGGGT
             520       530       540       550       560       570

    850       860       870       880       890       900
LEX151 CTCACTGCTGACTTCCTGGGGCGAGAGGCCATGATCTTCCGAAGTGGAGGTCCTCGGCCA
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 CTCACTGCTGACTTCCTGGGGCGAGAGGCCATGATCTTCCGAAGTGGAGGTCCTCGGCCA
             580       590       600       610       620       630

    910       920       930       940       950       960
LEX151 GCTCTGCGTTCCGACTCTGACCAGAGTCTCTTGCACGACCCCCGGTTTGTGATGGCCGCC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GCTCTGCGTTCCGACTCTGACCAGAGTCTCTTGCACGACCCCCGGTTTGTGATGGCCGCC
             640       650       660       670       680       690

    970       980       990      1000      1010      1020
LEX151 CGGATCCCTGAGAACTCTGACCAGGACAATGACAAGGTGTACTTCTTCTTCTCGGAGACG
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 CGGATCCCTGAGAACTCTGACCAGGACAATGACAAGGTGTACTTCTTCTTCTCGGAGACG
             700       710       720       730       740       750

   1030      1040      1050      1060      1070      1080
LEX151 GTCCCCTCGCCCGATGGTGGCTCGAACCATGTCACTGTCAGCCGCGTGGGCCGCGTCTGC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GTCCCCTCGCCCGATGGTGGCTCGAACCATGTCACTGTCAGCCGCGTGGGCCGCGTCTGC
             760       770       780       790       800       810

   1090      1100      1110      1120      1130      1140
LEX151 GTGAATGATGCTGGGGGCCAGCGGGTGCTGGTGAACAAATGGAGCACTTTCCTCAAGGCC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GTGAATGATGCTGGGGGCCAGCGGGTGCTGGTGAACAAATGGAGCACTTTCCTCAAGGCC
             820       830       840       850       860       870

   1150      1160      1170      1180      1190      1200
LEX151 AGGCTGGTCTGCTCGGTGCCCGGCCCTGGTGGTGCCGAGACCCACTTTGACCAGCTAGAG
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 AGGCTGGTCTGCTCGGTGCCCGGCCCTGGTGGTGCCGAGACCCACTTTGACCAGCTAGAG
             880       890       900       910       920       930

   1210      1220      1230      1240      1250      1260
```

```
LEX151 GATGTGTTCCTGCTGTGGCCCAAGGCCGGGAAGAGCCTCGAGGTGTACGCGCTGTTCAGC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GATGTGTTCCTGCTGTGGCCCAAGGCCGGGAAGAGCCTCGAGGTGTACGCGCTGTTCAGC
             940       950       960       970       980       990

    1270      1280      1290      1300      1310      1320
LEX151 ACCGTCAGTGCCGTGTTCCAGGGCTTCGCCGTCTGTGTGTACCACATGGCAGACATCTGG
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 ACCGTCAGTGCCGTGTTCCAGGGCTTCGCCGTCTGTGTGTACCACATGGCAGACATCTGG
            1000      1010      1020      1030      1040      1050

    1330      1340      1350      1360      1370      1380
LEX151 GAGGTTTTCAACGGGCCCTTTGCCCACCGAGATGGGCCTCAGCACCAGTGGGGGCCCTAT
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GAGGTTTTCAACGGGCCCTTTGCCCACCGAGATGGGCCTCAGCACCAGTGGGGGCCCTAT
            1060      1070      1080      1090      1100      1110

    1390      1400      1410      1420      1430      1440
LEX151 GGGGGCAAGGTGCCCTTCCCTCGCCCTGGCGTGTGCCCCAGCAAGATGACCGCACAGCCA
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GGGGGCAAGGTGCCCTTCCCTCGCCCTGGCGTGTGCCCCAGCAAGATGACCGCACAGCCA
            1120      1130      1140      1150      1160      1170

    1450      1460      1470      1480      1490      1500
LEX151 GGACGGCCTTTTGGCAGCACCAAGGACTACCCAGATGAGGTGCTGCAGTTTGCCCGAGCC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GGACGGCCTTTTGGCAGCACCAAGGACTACCCAGATGAGGTGCTGCAGTTTGCCCGAGCC
            1180      1190      1200      1210      1220      1230

    1510      1520      1530      1540      1550      1560
LEX151 CACCCCCTCATGTTCTGGCCTGTGCGGCCTCGACATGGCCGCCCTGTCCTTGTCAAGACC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 CACCCCCTCATGTTCTGGCCTGTGCGGCCTCGACATGGCCGCCCTGTCCTTGTCAAGACC
            1240      1250      1260      1270      1280      1290

    1570      1580      1590      1600      1610      1620
LEX151 CACCTGGCCCAGCAGCTACACCAGATCGTGGTGGACCGCGTGGAGGCAGAGGATGGGACC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 CACCTGGCCCAGCAGCTACACCAGATCGTGGTGGACCGCGTGGAGGCAGAGGATGGGACC
            1300      1310      1320      1330      1340      1350

    1630      1640      1650      1660      1670      1680
LEX151 TACGATGTCATTTTCCTGGGGACTGACTCAGGGTCTGTGCTCAAAGTCATCGCTCTCCAG
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 TACGATGTCATTTTCCTGGGGACTGACTCAGGGTCTGTGCTCAAAGTCATCGCTCTCCAG
            1360      1370      1380      1390      1400      1410

    1690      1700      1710      1720      1730      1740
LEX151 GCAGGGGGCTCAGCTGAACCTGAGGAAGTGGTTCTGGAGGAGCTCCAGGTGTTTAAGGTG
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GCAGGGGGCTCAGCTGAACCTGAGGAAGTGGTTCTGGAGGAGCTCCAGGTGTTTAAGGTG
            1420      1430      1440      1450      1460      1470

    1750      1760      1770      1780      1790      1800
LEX151 CCAACACCTATCACCGAAATGGAGATCTCTGTCAAAAGGCAAATGCTATACGTGGGCTCT
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 CCAACACCTATCACCGAAATGGAGATCTCTGTCAAAAGGCAAATGCTATACGTGGGCTCT
            1480      1490      1500      1510      1520      1530

    1810      1820      1830      1840      1850      1860
```

```
LEX151 CGGCTGGGTGTGGCCCAGCTGCGGCTGCACCAATGTGAGACTTACGGCACTGCCTGTGCA
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 CGGCTGGGTGTGGCCCAGCTGCGGCTGCACCAATGTGAGACTTACGGCACTGCCTGTGCA
            1540       1550       1560       1570       1580       1590

    1870       1880       1890       1900       1910       1920
LEX151 GAGTGCTGCCTGGCCCGGGACCCATACTGTGCCTGGGATGGTGCCTCCTGTACCCACTAC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GAGTGCTGCCTGGCCCGGGACCCATACTGTGCCTGGGATGGTGCCTCCTGTACCCACTAC
            1600       1610       1620       1630       1640       1650

    1930       1940       1950       1960       1970       1980
LEX151 CGCCCCAGCCTTGGCAAGCGCCGGTTCCGCCGGCAGGACATCCGGCACGGCAACCCTGCC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 CGCCCCAGCCTTGGCAAGCGCCGGTTCCGCCGGCAGGACATCCGGCACGGCAACCCTGCC
            1660       1670       1680       1690       1700       1710

    1990       2000       2010       2020       2030       2040
LEX151 CTGCAGTGCCTGGGCCAGAGCCAGGAAGAAGAGGCAGTGGGACTTGTGGCAGCCACCATG
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 CTGCAGTGCCTGGGCCAGAGCCAGGAAGAAGAGGCAGTGGGACTTGTGGCAGCCACCATG
            1720       1730       1740       1750       1760       1770

    2050       2060       2070       2080       2090       2100
LEX151 GTCTACGGCACGGAGCACAATAGCACCTTCCTGGAGTGCCTGCCCAAGTCTCCCCARGCT
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::.:::
gi|897 GTCTACGGCACGGAGCACAATAGCACCTTCCTGGAGTGCCTGCCCAAGTCTCCCCAGGCT
            1780       1790       1800       1810       1820       1830

    2110       2120       2130       2140       2150       2160
LEX151 GCTGTGCGCTGGCTCTTGCAGAGGCCAGGGGATGAGGGGCCTGACCAGGTGAAGACGGAC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GCTGTGCGCTGGCTCTTGCAGAGGCCAGGGGATGAGGGGCCTGACCAGGTGAAGACGGAC
            1840       1850       1860       1870       1880       1890

    2170       2180       2190       2200       2210       2220
LEX151 GAGCGAGTCTTGCACACGGAGCGGGGGCTGCTGTTCCGCAGGCTTAGCCGTTTCGATGCG
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GAGCGAGTCTTGCACACGGAGCGGGGGCTGCTGTTCCGCAGGCTTAGCCGTTTCGATGCG
            1900       1910       1920       1930       1940       1950

    2230       2240       2250       2260       2270       2280
LEX151 GGCACCTACACCTGCACCACTCTGGAGCATGGCTTCTCCCAGACTGTGGTCCGCCTGGCT
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GGCACCTACACCTGCACCACTCTGGAGCATGGCTTCTCCCAGACTGTGGTCCGCCTGGCT
            1960       1970       1980       1990       2000       2010

    2290       2300       2310       2320       2330       2340
LEX151 CTGGTGGTGATTGTGGCCTCACAGCTGGACAACCTGTTCCCTCCGGAGCCAAAGCCAGAG
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 CTGGTGGTGATTGTGGCCTCACAGCTGGACAACCTGTTCCCTCCGGAGCCAAAGCCAGAG
            2020       2030       2040       2050       2060       2070

    2350       2360       2370       2380       2390       2400
LEX151 GAGCCCCCAGCCCGGGGAGGCCTGGCTTCCACCCCACCCAAGGCCTGGTACAAGGACATC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 GAGCCCCCAGCCCGGGGAGGCCTGGCTTCCACCCCACCCAAGGCCTGGTACAAGGACATC
            2080       2090       2100       2110       2120       2130

    2410       2420       2430       2440       2450       2460
```

```
LEX151 CTGCAGCTCATTGGCTTCGCCAACCTGCCCCGGGTGGATGAGTACTGTGAGCGCGTGTGG
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 CTGCAGCTCATTGGCTTCGCCAACCTGCCCCGGGTGGATGAGTACTGTGAGCGCGTGTGG
            2140      2150      2160      2170      2180      2190

    2470      2480      2490      2500      2510      2520
LEX151 TGCAGGGGCACCACGGAATGCTCAGGCTGCTTCCGGAGCCGGAGCCGGGGCAAGCAGGCC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 TGCAGGGGCACCACGGAATGCTCAGGCTGCTTCCGGAGCCGGAGCCGGGGCAAGCAGGCC
            2200      2210      2220      2230      2240      2250

    2530      2540      2550      2560      2570      2580
LEX151 AGGGGCAAGAGCTGGGCAGGGCTGGAGCTAGGCAAGAAGATGAAGAGCCGGGTGCATGCC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|897 AGGGGCAAGAGCTGGGCAGGGCTGGAGCTAGGCAAGAAGATGAAGAGCCGGGTGCATGCC
            2260      2270      2280      2290      2300      2310

    2590      2600      2610      2620
LEX151 GAGCACAATCGGACGCCCCGGGAGGTGGAGGCCACGTAG
       ::::::::::::::::::::::::::::::::::::::
gi|897 GAGCACAATCGGACGCCCCGGGAGGTGGAGGCCACGTAGAAGGGGGCAGAGGAGGGGTGG
            2320      2330      2340      2350      2360      2370

gi|897 TCAGGATGGGCTGGGGGGCCCACTAGCAGCCCCCAGCATCTCCCACCCACCCAGCTAGGG
            2380      2390      2400      2410      2420      2430

>>gi|8978201|dbj|AB029496.1| Homo sapiens mRNA for semap  (4700 nt)
rev-comp initn:  83 init1:  83 opt:  95
 67.105% identity in 76 nt overlap (119-46:407-475)

    150       140       130       120       110       100
LEX15- GCGGGAAGAGGGGCGGAGGAGAGAAGGAGGCTGGGGCCTTGCCGTCCACCTGCCGCTTCT
                                  :::::::::: : : :::::::   :  ::
gi|897 ACAACCGGACCCACCTGCTAGCCTGTGGCACTGGGGCCTTCCAGCCCACCTGTGCC--CT
        380       390       400       410       420       430

     90        80         70         60        50        40
LEX15- CCTTCCACCTTGTTGGCC-CAGTGCAG-GCTTTTGTGCCACACTGGCCAGCTCCCCATTG
       : :  ::    ::::::: : :::  : :: : ::  :::: ::::
gi|897 CATCACA----GTTGGCCACCGTGGGGAGCATGTGCTCCAC-CTGGAGCCTGGCAGTGTG
          440           450       460       470       480

       30        20        10
LEX15- GGAAGACCTTCCCAGCTAGGGCACAGGCCAT

gi|897 GAAAGTGGCCGGGGGCGGTGCCCTCACGAGCCCAGCCGTCCCTTTGCCAGCACCTTCATA
     490       500       510       520       530       540


2628 residues in 1 query  sequences
4700 residues in 1 library sequences
 Scomplib [version 3.3t05 March 30, 2000]
 start: Fri Sep 19 13:50:44 2003 done: Fri Sep 19 13:50:45 2003
 Scan time:  0.117 Display time:  0.133

Function used was FASTA
```

```
 FASTA searches a protein or DNA sequence data bank
 version 3.3t05 March 30, 2000
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

/tmp/fastaGAAvaaG8P: 781 aa
>SEQ ID NO:3 human semaphorin
 vs  /tmp/fastaHAAwaaG8P library
searching /tmp/fastaHAAwaaG8P library

   814 residues in     1 sequences

FASTA (3.34 January 2000) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 38, opt: 26, gap-pen: -12/ -2, width:  16
 Scan time:  0.034
The best scores are:                                          opt
SEQ ID NO:1 human semaphorin MACALAGKVFPMGSWPVWHK  ( 814) 5462

>>SEQ ID NO:1 human semaphorin MACALAGKVFPMGSWPVWHKSLHWA  (814 aa)
 initn: 5462 init1: 5462 opt: 5462
Smith-Waterman score: 5462;  100.000% identity in 781 aa overlap (1-781:34-814)

                                        10        20        30
SEQ                                    MAPSAWAICWLLGGLLLHGGSSGPSPGPSV
                                       ::::::::::::::::::::::::::::::
SEQ    GGSRANYNRRPAGPEGGSAGRRQRCPQFPSMAPSAWAICWLLGGLLLHGGSSGPSPGPSV
            10        20        30        40        50        60

             40        50        60        70        80        90
SEQ    PRLRLSYRDLLSANRSAIFLGPQGSLNLQAMYLDEYRDRLFLGGLDALYSLRLDQAWPDP
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
SEQ    PRLRLSYRDLLSANRSAIFLGPQGSLNLQAMYLDEYRDRLFLGGLDALYSLRLDQAWPDP
            70        80        90       100       110       120

            100       110       120       130       140       150
SEQ    REVLWPPQPGQREECVRKGRDPLTECANFVRVLQPHNRTHLLACGTGAFQPTCALITVGH
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
SEQ    REVLWPPQPGQREECVRKGRDPLTECANFVRVLQPHNRTHLLACGTGAFQPTCALITVGH
           130       140       150       160       170       180

            160       170       180       190       200       210
SEQ    RGEHVLHLEPGSVESGRGRCPHEPSRPFASTFIDGELYTGLTADFLGREAMIFRSGGPRP
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
SEQ    RGEHVLHLEPGSVESGRGRCPHEPSRPFASTFIDGELYTGLTADFLGREAMIFRSGGPRP
           190       200       210       220       230       240

            220       230       240       250       260       270
SEQ    ALRSDSDQSLLHDPRFVMAARIPENSDQDNDKVYFFFSETVPSPDGGSNHVTVSRVGRVC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
SEQ    ALRSDSDQSLLHDPRFVMAARIPENSDQDNDKVYFFFSETVPSPDGGSNHVTVSRVGRVC
           250       260       270       280       290       300

            280       290       300       310       320       330
SEQ    VNDAGGQRVLVNKWSTFLKARLVCSVPGPGGAETHFDQLEDVFLLWPKAGKSLEVYALFS
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
SEQ    VNDAGGQRVLVNKWSTFLKARLVCSVPGPGGAETHFDQLEDVFLLWPKAGKSLEVYALFS
           310       320       330       340       350       360

            340       350       360       370       380       390
SEQ    TVSAVFQGFAVCVYHMADIWEVFNGPFAHRDGPQHQWGPYGGKVPFPRPGVCPSKMTAQP
```

```
          :::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
SEQ       TVSAVFQGFAVCVYHMADIWEVFNGPFAHRDGPQHQWGPYGGKVPFPRPGVCPSKMTAQP
              370       380       390       400       410       420

                400       410       420       430       440       450
SEQ       GRPFGSTKDYPDEVLQFARAHPLMFWPVRPRHGRPVLVKTHLAQQLHQIVVDRVEAEDGT
          :::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
SEQ       GRPFGSTKDYPDEVLQFARAHPLMFWPVRPRHGRPVLVKTHLAQQLHQIVVDRVEAEDGT
              430       440       450       460       470       480

                460       470       480       490       500       510
SEQ       YDVIFLGTDSGSVLKVIALQAGGSAEPEEVVLEELQVFKVPTPITEMEISVKRQMLYVGS
          :::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
SEQ       YDVIFLGTDSGSVLKVIALQAGGSAEPEEVVLEELQVFKVPTPITEMEISVKRQMLYVGS
              490       500       510       520       530       540

                520       530       540       550       560       570
SEQ       RLGVAQLRLHQCETYGTACAECCLARDPYCAWDGASCTHYRPSLGKRRFRRQDIRHGNPA
          :::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
SEQ       RLGVAQLRLHQCETYGTACAECCLARDPYCAWDGASCTHYRPSLGKRRFRRQDIRHGNPA
              550       560       570       580       590       600

                580       590       600       610       620       630
SEQ       LQCLGQSQEEEAVGLVAATMVYGTEHNSTFLECLPKSPAAVRWLLQRPGDEGPDQVKTDE
          :::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
SEQ       LQCLGQSQEEEAVGLVAATMVYGTEHNSTFLECLPKSPAAVRWLLQRPGDEGPDQVKTDE
              610       620       630       640       650       660

                640       650       660       670       680       690
SEQ       RVLHTERGLLFRRLSRFDAGTYTCTTLEHGFSQTVVRLALVVIVASQLDNLFPPEPKPEE
          :::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
SEQ       RVLHTERGLLFRRLSRFDAGTYTCTTLEHGFSQTVVRLALVVIVASQLDNLFPPEPKPEE
              670       680       690       700       710       720

                700       710       720       730       740       750
SEQ       PPARGGLASTPPKAWYKDILQLIGFANLPRVDEYCERVWCRGTTECSGCFRSRSRGKQAR
          :::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
SEQ       PPARGGLASTPPKAWYKDILQLIGFANLPRVDEYCERVWCRGTTECSGCFRSRSRGKQAR
              730       740       750       760       770       780

                760       770       780
SEQ       GKSWAGLELGKKMKSRVHAEHNRTPREVEAT
          ::::::::::::::::::::::::::::::
SEQ       GKSWAGLELGKKMKSRVHAEHNRTPREVEAT
              790       800       810
```

781 residues in 1 query sequences
814 residues in 1 library sequences
Scomplib [version 3.3t05 March 30, 2000]
start: Mon Feb 10 16:26:11 2003 done: Mon Feb 10 16:26:12 2003
Scan time: 0.034 Display time: 0.866

Function used was FASTA

12. S. Ohno, *Curr. Opin. Cell Biol.* 13, 641 (2001).
13. D. Bilder *et al.*, *Nature Cell Biol.* 2, E114 (2000).
14. D. Bilder, M. Li, N. Perrimon, *Science* 289, 113 (2000).
15. D. Bilder, M. Schober, N. Perrimon, *Nature Cell Biol.*, in press.
16. S. Baumgartner *et al.*, *Cell* 87, 1059 (1996).
17. R. E. Ward, R. S. Lamb, R. G. Fehon, *J. Cell Biol.* 140, 1463 (1998).
18. D. F. Woods, P. J. Bryant, *Cell* 66, 451 (1991).
19. J. S. Simske, J. Hardin, *BioEssays* 23, 12 (2001).
20. M. Köppen *et al.*, *Nature Cell Biol.* 3, 983 (2001).
21. B. Leung, G. J. Hermann, J. R. Priess, *Dev. Biol.* 216, 114 (1999).
22. L. McMahon, R. Legouis, J. L. Vonesch, M. Labouesse, *J. Cell Sci.* 114, 2265 (2001).
23. J. E. Sulston, E. Schierenberg, J. G. White, J. N. Thomson, *Dev. Biol.* 100, 64 (1983).
24. M. F. Maduro, J. H. Rothman, *Dev. Biol.* 246, 68 (2002).
25. O. Bossinger, A. Klebes, C. Segbert, C. Theres, E. Knust, *Dev. Biol.* 230, 29 (2001).
26. B. Podbilewicz, J. G. White, *Dev. Biol.* 161, 408 (1994).
27. M. Labouesse, *Dev. Dyn.* 210, 19 (1997).
28. R. Legouis *et al.*, *Nature Cell Biol.* 2, 415 (2000).
29. J. Pellettieri, G. Seydoux, *Science* 298, 1946 (2002).
30. D. D. Hurd, K. Kemphues, *Dev. Biol.*, in press.
31. B. L. Firestein, C. Rongo, *Mol. Biol. Cell* 12, 3465 (2001).
32. H. Hutter *et al.*, *Science* 287, 989 (2000).
33. L. Chen, B. Ong, V. Bennett, *J. Cell Biol.* 154, 841 (2001).
34. Y. Watari *et al.*, *Gene* 224, 53 (1998).
35. N. Kioka, K. Ueda, T. Amachi, *Cell Struct. Funct.* 27, 1 (2002).
36. S. Tsukita, M. Furuse, M. Itoh, *Nature Rev. Mol. Cell. Biol.* 2, 285 (2001).
37. M. H. Roh *et al.*, *J. Cell Biol.* 157, 161 (2002).
38. M. Furuse *et al.*, *J. Cell Biol.* 156, 1099 (2002).
39. M. Furuse, H. Sasaki, K. Fujimoto, S. Tsukita, *J. Cell Biol.* 143, 391 (1998).
40. K. Ebnet *et al.*, *EMBO J.* 20, 3738 (2001).
41. M. Itoh *et al.*, *J. Cell Biol.* 154, 491 (2001).
42. C. Lemmers *et al.*, *J. Biol. Chem.* 277, 25408 (2002).
43. M. H. Roh, C. J. Liu, S. Laurinec, B. Margolis, *J. Biol. Chem.* 277, 27501 (2002).
44. A. I. den Hollander *et al.*, *Mech. Dev.* 110, 203 (2002).
45. A. I. den Hollander *et al.*, *Am. J. Hum. Genet.* 69, 198 (2001).
46. M. Pellikka *et al.*, *Nature* 416, 143 (2002).
47. S. Izaddoost, S. C. Nam, M. A. Bhat, H. J. Bellen, K. W. Choi, *Nature* 416, 178 (2002).
48. K. Johnson, F. Grawe, N. Grzeschik, E. Knust, *Curr. Biol.* 12, 1675 (2002).
49. A. Wodarz, *Nature Cell Biol.* 4, E39 (2002).
50. We thank B. Margolis for providing Fig. 1E and P. Hardy for critical reading of the manuscript. We apologize to colleagues whose work is not described or referenced here because of space limitations. Work in the laboratories of E.K. and O.B. is supported by grants from the Deutsche Forschungsgemeinschaft and the Fonds der Chemischen Industrie.

REVIEW

Molecular Mechanisms of Axon Guidance

Barry J. Dickson

Axons are guided along specific pathways by attractive and repulsive cues in the extracellular environment. Genetic and biochemical studies have led to the identification of highly conserved families of guidance molecules, including netrins, Slits, semaphorins, and ephrins. Guidance cues steer axons by regulating cytoskeletal dynamics in the growth cone through signaling pathways that are still only poorly understood. Elaborate regulatory mechanisms ensure that a given cue elicits the right response from the right axons at the right time but is otherwise ignored. With such regulatory mechanisms in place, a relatively small number of guidance factors can be used to generate intricate patterns of neuronal wiring.

The correct wiring of the nervous system relies on the uncanny ability of axons and dendrites to locate and recognize their appropriate synaptic partners. To help them find their way in the developing embryo, axons and dendrites are tipped with a highly motile and exquisitely sensitive structure, the growth cone. Extracellular guidance cues can either attract or repel growth cones, and can operate either at close range or over a distance (*1*). By responding to the appropriate set of cues, growth cones are able to select the correct path toward their target.

Ten years ago (*2*), very few of the molecules that guide axons in vivo were known. But the 1970s and '80s had seen the introduction of several powerful in vitro assays to detect guidance activities in the developing vertebrate nervous system, and the growing interest of invertebrate geneticists in the problem of axon guidance. So by the early 1990s, the stage had been set for a burst of activity that led to the discovery of several conserved families of axon guidance molecules. Prominent among these are the netrins, Slits, semaphorins, and ephrins (Fig. 1).

Research Institute of Molecular Pathology, Dr. Bohrgasse 7, A-1030 Vienna, Austria. E-mail: dickson@nt.imp.univie.ac.at

These are not the only known guidance molecules, but they are by far the best understood. With these molecules in hand, we can now begin to ask how growth cones sense and respond to guidance cues, and how a relatively small number of cues can be used to assemble complex neuronal networks.

Guidance Cues and Their Receptors

Netrins. The discovery of netrins came as the remarkable convergence of the search for a chemoattractant for vertebrate commissural axons (*3*, *4*), and the analysis of genes required for circumferential axon guidance in *Caenorhabditis elegans* (*5*, *6*). Across more than 600 million years of evolution, netrins have retained the function of attracting axons ventrally toward the midline (*7*). Netrins can also repel some axons, and this function too has been conserved. This was initially inferred from defects in dorsal as well as ventral guidance in *unc-6/netrin* mutant worms (*5*), and subsequently confirmed by the direct demonstration of netrin's repulsive activity in vertebrates (*8*) and in flies (*9*, *10*).

Identification of the netrin receptors followed from the characterization of two other worm mutants with defects in circumferential guidance: *unc-40*, which primarily disrupts ventral guidance; and *unc-5*, which affects only dorsal guidance (*5*). Both *unc-40* and *unc-5* encode conserved transmembrane proteins (*7*), with UNC-40 belonging to the DCC (deleted in colorectal carcinoma) family. Biochemical and genetic studies have confirmed their functions as netrin receptors in several different species (*7*, *10*). DCC receptors mediate attraction to netrins but can also participate in repulsion. UNC-5 receptors appear to function exclusively in repulsion, either alone or in combination with DCC receptors. UNC-5 receptors may require a DCC coreceptor for repulsion farther away from the netrin source, where ligand concentration is likely to be lower (*5*, *10*). This may involve a direct interaction between the cytoplasmic domains of the two receptors (*11*).

Netrins guide many different axons in vivo. In some cases, netrin can exert its effects from distances of up to a few millimeters (*12*), but in others it appears to act only at short range (*9*). Netrins have high affinity for cell membranes (*3*, *4*), and it is unclear how far they can diffuse in vivo and how their diffusion is regulated. Indeed, a netrin gradient has not yet been visualized directly in any system, and formal proof that netrin must diffuse away from its source to exert its long-range effects is lacking.

Slits. Slits are large secreted proteins that signal through Roundabout (Robo) family receptors. Robo was first identified in a genetic screen for midline guidance defects in *Drosophila* (*13*, *14*). Genetic studies suggested that Robo is the receptor for a midline repellent (*14*), subsequently identified as Slit (*15*, *16*). This repulsive action of Slit was found to be conserved in vertebrates (*17*, *18*). However, in a parallel approach, Slit was also purified as a factor that stimulates sensory axon branching

and elongation (*19*). Thus, Slits, like netrins, are multifunctional. The importance of Slit-Robo signaling in axon guidance and cell migration is underscored by the recovery of *robo* mutations in genetic screens for guidance defects in at least three different species (*13*, *20*, *21*) and the purification of Slit in at least three independent biochemical assays (*19*, *22*, *23*).

The best-understood functions of Slit proteins are in midline guidance in *Drosophila* and in the formation of the optic chiasm in vertebrates. In *Drosophila*, Slit is expressed at the ventral midline, where it acts as a short-range repellent signaling through Robo to prevent ipsilateral axons from crossing the midline and commissural axons from recrossing (*15*, *16*). Two other Slit receptors, Robo2 and Robo3, specify the lateral positions of axons that run parallel to the midline, presumably in response to a long-range gradient of Slit activity diffusing away from the midline (*24*, *25*).

Vertebrate Slit proteins are also expressed by ventral midline cells (*17*), and commissural axons are repelled by Slit after they have crossed the midline (*26*). Mice deficient for both Slit1 and Slit2 lack any obvious defects in midline guidance in the spinal cord (*27*), but Slit3 is still expressed at the midline in these mice.

Slit1/2-deficient mice do have striking defects in the formation of the optic chiasm, where Slit3 is not expressed (*27*). These defects are strikingly reminiscent of those seen in *astray/robo2* mutant fish, in which retinal axons make multiple guidance errors before, during, and after crossing the midline (*21*). Similar errors also occur in wild-type fish but are always corrected (*28*). In fish, all retinal axons project contralaterally, but in mice, which have binocular vision, some axons project contralaterally and others ipsilaterally. By analogy to the role of *Drosophila* Slit in midline guidance, it was anticipated that the vertebrate Slit proteins might be expressed at the chiasm and control the choice of an ipsilateral or contralateral projection. This is not the case (*29*). Instead, Slit1 and Slit2 are expressed by cells surrounding the chiasm and repel ipsilateral and contralateral axons alike (*23*, *27*, *30*, *31*). This has led to the idea that Slits form a repulsive corridor to guide all retinal axons through the chiasm.

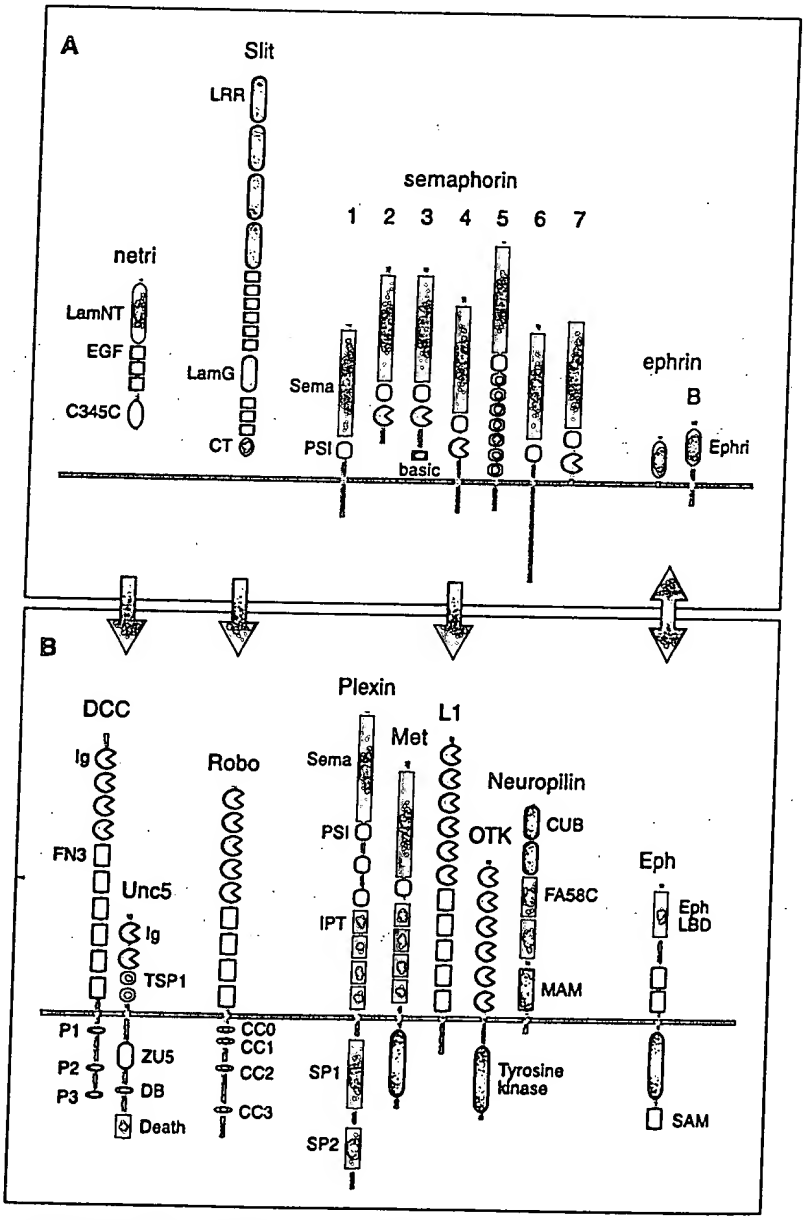


Fig. 1. Conserved families of guidance molecules (A) and their receptors (B). Domain names are from SMART (http://smart.embl-heidelberg.de). P1 to P3, DB (DCC-binding), CC0 to CC3, and SP1 and SP2 indicate conserved regions in the cytoplasmic domains of DCC, UNC-5, Robo, and Plexin receptors, respectively.

Semaphorins. Semaphorins are a large family of cell surface and secreted guidance molecules, defined by the presence of a conserved ~420–amino acid Sema domain at their NH_2-termini. The first semaphorins were identified by searching for molecules expressed on specific axon fascicles in the grasshopper central nervous system (CNS) (*32*) and by purifying a potent inducer of vertebrate sensory growth cone collapse in vitro (*33*). Semaphorins are divided into eight classes, on the basis of their structure. Classes 1 and 2 are found in invertebrates, classes 3 to 7 are found in vertebrates, and class V semaphorins are encoded by viruses (*34*).

Semaphorins signal through multimeric receptor complexes. The composition of these receptor complexes is not fully known. Many, and perhaps all, semaphorin receptor complexes include a plexin protein. Plexins comprise a large family of transmembrane proteins divided into four groups (A to D), on the basis of sequence similarity (*35*). *Drosophila* PlexinA is a functional receptor for the transmembrane Sema1a (*36*), vertebrate plexin-As are functional receptors for secreted class 3 semaphorins (*35*, *37*), and other plexins bind directly to semaphorins of different classes (*35*, *38*, *39*). Receptor complexes for the vertebrate class 3 semaphorins also include neuropilins, which bind directly to both semaphorins and plexins (*34*). Neuropilins do not appear to have a signaling function, but rather contribute to ligand specificity. Other essential components of semaphorin receptor complexes include the neural cell adhesion molecule L1 (for Sema3A) (*40*), the receptor tyrosine kinase Met (for Sema4D) (*41*), and the catalytically inactive receptor tyrosine kinase OTK (for *Drosophila* Sema1a) (*42*).

Genetic analysis of semaphorin function in flies and in mice suggests that they primarily act as short-range inhibitory cues that deflect axons away from inappropriate re-

gions, or guide them through repulsive corridors (*34*, *37*). Evidence suggests that semaphorins may also act as attractive cues for certain axons (*34*, *43*), although this remains to be verified by genetic analysis. Interestingly, semaphorins do not seem to function in axon guidance in *C. elegans*, but instead have an analogous role in discouraging inappropriate cell contacts. Worms have three semaphorin and two plexin genes, all of which have been mutated (*44*–*46*). In these mutants, epidermal cells that should only transiently contact one another instead make more perdurant contacts.

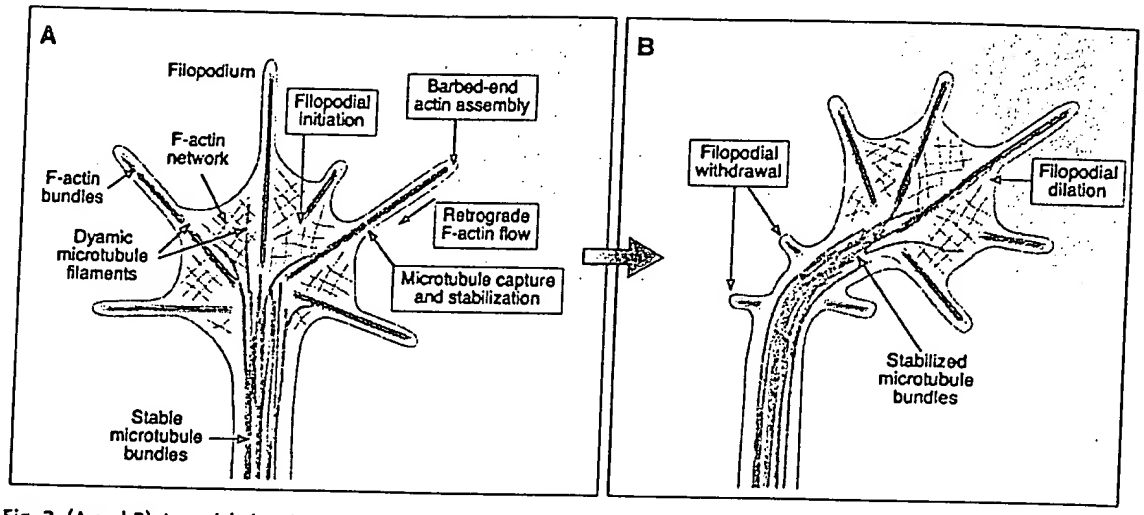


Fig. 2. (A and B) A model showing one way in which a growth cone might turn toward an attractant (green).

Ephrins. In a classic paper (*47*), Sperry postulated that vertebrate retinal axons are guided to their appropriate topographic locations in the optic tectum by an orthogonal system of molecular gradients in the retina and the tectum. The search for these graded cues led to the identification of the ephrins, membrane-bound ligands for the Eph family of receptor tyrosine kinases (*48*, *49*). Ephrins and Eph receptors fall into two classes: ephrin-As, which are anchored to the membrane by a glycosylphosphatidylinositol (GPI) linkage and bind EphA receptors; and ephrin-Bs, which have a transmembrane domain and bind EphB receptors (*50*).

In the visual system, topographic mapping of retinal axons along the anterior-posterior axis depends on repulsion mediated by ephrin-A ligands and their EphA receptors (*50*). Ephrin-A ligands are expressed in a gradient in the tectum [or its mammalian equivalent, the superior colliculus (SC)], and EphA receptors are expressed in a complementary gradient in the retina. Retinal axons with successively higher EphA levels map to successively lower points along the ephrin-A gradient. If the ephrin-A gradient is eliminated in the mouse SC, then retinal axons do not all shift to one end of the SC, as would be expected if each retinal axon simply mapped to a specific threshold value on the ephrin-A gradient. Instead, retinal axons still fill the entire SC, but their topographic order is disrupted—some axons shift posteriorly and others anteriorly (*51*). This suggests that the ephrin-A gradient establishes the topographic order of retinal axons, but not their precise termination sites. Further support for this model comes from a clever genetic experiment in which half the retinal axons were forced to express higher levels of an EphA receptor (*52*). Those axons with extra EphA receptors shifted down the ephrin-A gradient, whereas those with only their endogenous levels shifted up the gradient. The result was two smooth maps, one in each half of the SC. The conclusion is that the mapping of retinal axons depends on their relative EphA levels, not their absolute levels.

Mapping along the dorsal-ventral axis, in contrast, involves attractive signaling mediated by ephrin-B ligands and EphB receptors (*53*, *54*). Correct mapping of retinal axons along this axis evidently requires both "forward" signaling, in which ephrin-B ligands activate EphB receptors, and "reverse" signaling, in which EphBs serve as ligands to signal back through the transmembrane ephrin-Bs.

Ephrins control axon guidance in many other places too, and the ability to signal in either direction is a common theme, as is the ability to mediate either attraction or repulsion (*50*). For example, ephrin-B reverse signaling repels forebrain commissural axons away from regions of EphB expression (*55*) while attracting them to regions of EphA4 expression (*56*). The GPI-anchored ephrin-As are also able to signal in the reverse direction (*57*) and may act in this mode to mediate attraction or adhesion during mapping of vomeronasal axons to the accessory olfactory bulb (*58*).

Mammals have 13 Eph receptors and 8 ephrins. Worms and flies both have just a single Eph receptor, with four and one ephrin ligands, respectively. Somewhat surprisingly, the invertebrate ephrin and Eph mutants do not have dramatic axon guidance defects (*59*–*63*). The *C. elegans* ephrins and the Eph receptor do, however, have critical functions in multiple aspects of epithelial morphogenesis, as do their vertebrate counterparts (*50*). It seems that ephrins and Eph receptors are an ancient but versatile system for cell-cell communication that has diversified and acquired its axon guidance functions primarily during vertebrate evolution.

Steering the Growth Cone

Cytoskeleton. Growth cone turning is a complex process in which actin-based motility is harnessed to produce persistent and directed microtubule advance (Fig. 2). Actin filaments are organized into two distinct populations: dense, parallel filaments that radiate outward and into filopodia; and intervening networks of loosely interwoven filaments (*64*). Filopodial filaments are oriented with their fast-growing barbed ends toward the filopodium tip. The extension and retraction of a filopodium reflect the balance between the polymerization of actin at barbed ends and the retrograde flow of entire filaments (*65*–*67*). Filopodia often extend asymmetrically before the entire growth cone turns (*68*–*70*), and without filopodia, growth cones become disoriented (*69*, *71*, *72*). The precise role of filopodia in growth cone turning remains unclear, but they have been postulated to steer the growth cone by differential adhesion (*73*), generating mechanical force (*74*), or transducing distal signals (*75*).

Microtubules form stable, cross-linked bundles in the axon shaft. Single microtubule filaments also emerge into the growth cone. These filaments display the classic properties of dynamic instability, extending and retracting as they explore the peripheral region of the growth cone (*76*). These dynamic microtubules grow preferentially along the filopodial actin filaments (*76*, *77*), and the capture or stabilization of microtubule bundles in a specific filopodium may be a critical event in growth cone turning. Consistent with this view, stabilization and dilation of a single filopodium appear to be a common feature of growth cone turning in vivo (*78*–*80*).

There are many different ways in which a guidance signal might intervene to steer the growth cone. For example, a guidance cue might promote the initiation, extension, stabilization, or retraction of individual filopodia, or the capture or stabilization of microtubules in specific regions of the growth cone. Likely targets for the signaling pathways downstream of guidance receptors are therefore molecules such as Arp2/3 (to nucleate new actin filaments), Ena/VASP proteins (to promote filament elongation), adhesion molecules (to couple actin filaments to the substrate), and myosins (to regulate the retrograde flow of actin filaments). Molecules that capture microtubule ends (e.g., IQGAP1) or suppress microtubule instability (e.g., MAP1B) are also potential targets for guidance signals. We still need to determine which aspect(s) of actin or microtubule dynamics are the primary targets for regulation by each of the known guidance cues. It is difficult to trace the signaling pathways downstream of a guidance receptor without knowing what lies at the "business end."

Signaling. With their well-known roles in regulating cytoskeletal dynamics in fibroblasts, Rho guanosine triphosphatases (GTPases) were strong candidates to transduce guidance signals in the growth cone. A function for Rho GTPases in growth cone guidance was suggested from studies with dominant mutant isoforms (*81*) and was confirmed by the analysis of loss-of-function mutations in flies and worms (*82–85*). Biochemical links have also been made between several guidance receptors and Rho GTPases. For example, EphA receptors regulate the guanine nucleotide exchange factor (GEF) Ephexin (*86*); Robo receptors may act at least in part by regulating GTPase-activating proteins (GAPs) (*87*); and Plexins bind directly to Rho GTPases (*88*) and Rho GEFs (*39*), and may even have intrinsic GAP activity (*89*). Several downstream effectors of Rho GTPases have also been implicated in axon growth and guidance, such as Pak (*90*) and Rho kinase (*91*).

Genetic studies have also revealed important roles for Ena/VASP proteins in axon guidance (*92–94*). These proteins antagonize capping proteins to promote actin filament elongation (*95*). In motile fibroblasts, Ena/VASP proteins localize to the leading edge of lamellipodia. Depletion of Ena/VASP proteins from the leading edge leads to shorter, more highly branched filaments that generate greater protrusive force and increased motility. Conversely, increasing Ena/VASP levels at the leading edge results in longer, unbranched filaments and reduced motility (*95*, *96*). Genetic studies implicating Ena/VASP proteins in repulsive growth cone guidance by both Slit (*94*) and netrin (*92*) have been interpreted in light of this negative role in fibroblast motility. However, in growth cones, Ena/VASP proteins localize to filopodial tips (*97*), where actin filaments are normally unbranched and stable. Here, their activity would be expected to promote filopodial extension, making a role in attractive guidance equally plausible.

These considerations raise an important point. Migrating fibroblasts and axonal growth cones can have very different cytoskeletal organizations, and the location and action of molecules such as Ena/VASP proteins and Rho GTPases in growth cones cannot be inferred merely by analogy to fibroblasts. It will be important to determine precisely when, where, and how these proteins function in growth cones.

Calcium signaling may also play an important role in growth cone turning. In cultured *Xenopus* spinal neurons, turning in response to a netrin-1 gradient requires calcium influx through the plasma membrane, as well as calcium release from intracellular stores (*98*). Moreover, netrin-1 induces a transient Ca^{2+} gradient in the growth cone (*98*), and the creation of such a gradient by local photolysis of caged Ca^{2+} or release from intracellular stores is sufficient to induce turning in the absence of netrin-1 (*98*, *99*). Spontaneous calcium transients have also been observed in growth cones (*100*) and in filopodia (*101*). The frequencies of these transients appear to correlate negatively with growth cone extension rates, but compelling evidence of their involvement in growth cone turning in vivo is lacking.

Plasticity of Guidance Responses

Axons can evidently differ in their response to the same cue, as they must if they are to follow divergent pathways. But even a single growth cone may need to respond to the same cue in different ways at different points along its journey. This is particularly true if the growth cone is to navigate through a series of intermediate targets before reaching its final goal, as many do. Specifying an axon's trajectory is therefore not just a simple matter of selecting the appropriate set of guidance receptors and delivering them to the growth cone. The growth cone must also be able to modulate its responsiveness en route. Some of the mechanisms underlying this plasticity have recently come to light.

Modulation by cyclic nucleotides. In vitro, the responses of *Xenopus* spinal axons can be modulated by changing the levels of cyclic nucleotides (*102–104*). Responses to some guidance cues, including netrin-1, are sensitive to levels of cAMP or protein kinase A (PKA) activity, while others, including Sema3A, are modulated by cGMP and protein kinase G (PKG). The general finding is that lowering cAMP or cGMP levels or inhibiting PKA or PKG, converts an attractive response to a repulsive one, whereas elevating cAMP or cGMP, or activating PKA or PKG, switches repulsion to attraction.

Modulation of netrin-1 responsiveness by cAMP levels may play an important role in pathfinding of *Xenopus* retinal axons to the tectum (*105*). These axons are first attracted out of the eye by netrin-1 at the optic nerve head, become indifferent to it as they then grow through the ventral diencephalon, and finally are repelled by netrin-1 once they reach the tectum. These changes correlate with a gradual decline in cAMP levels and can be reversed by artificially raising cAMP levels. An intriguing variation on this theme has been documented in the mammalian cortex (*106*). Sema3A attracts the apical dendrites of pyramidal neurons toward the cortical plate but repels their axons away from it. Interestingly, a guanylyl cyclase is specifically localized in dendrites, implying that cGMP levels may be higher in dendrites than in axons.

Local translation in the growth cone. Applying netrin-1 or Sema3A to cultured *Xenopus* retinal axons induces local protein synthesis within the growth cone, and blocking translation inhibits the turning but not the growth of these axons (*107*). Induced protein synthesis is rapid enough to contribute directly to growth cone steering, but work on *Xenopus* spinal axons suggests a more subtle role: Growth cones might need to synthesize new proteins to maintain their sensitivity as they migrate up or down a ligand gradient (*108*). Spinal growth cones undergo consecutive phases of desensitization and resensitization to netrin-1 in vitro, and resensitization requires protein synthesis. Inhibiting translation in spinal axons does not block turning toward the netrin-1 source, as it does in retinal axons (*107*), but actually causes turning away from it (*108*). This is difficult to explain if translation has a direct role in growth cone turning, but could be explained by a role in resensitization: If desensitization is more rapid on the side of the growth cone facing toward the source, then a failure to synthesize the new proteins needed for resensitization could result, paradoxically, in a stronger attractive signal on the side facing away from the source.

Local translation might also be used to completely switch the growth cone's responsiveness to specific cues once it reaches an intermediate target. Evidence for such a mechanism comes from the finding that the 3′-untranslated region of the *EphA2* mRNA contains a sequence that confers selective translation in the distal segments of commissural axons, after they have crossed the midline (*109*). This could explain why the EphA2 receptor is only expressed at high levels in the segments of these axons that extend beyond the midline. The implication is that commissural axons might become sensitive to the ephrinA ligands in the spinal

cord only after crossing, although this remains to be tested.

Switching responses at the midline. To reach their targets on the contralateral side of the CNS, commissural axons must first grow toward the midline, but then leave it again on the opposite side and never turn back. Experiments in rodents, chicks, and flies have suggested a simple model for this behavior, in which commissural growth cones switch their sensitivity to midline attractants and repellents as they cross (Fig. 3). Before crossing, commissural axons are attracted to the midline by netrin (*3*, *4*) but are insensitive to the midline repellents Slit and, in vertebrates, certain class 3 semaphorins (*17*, *26*). After crossing, these axons are insensitive to netrins (at least in the vertebrate hindbrain) (*110*) but are repelled by both Slits and semaphorins (*26*). What turns attraction off and repulsion on at the midline?

One way in which netrin attraction could be turned off is by exposure to Slit. This is suggested by studies on cultured *Xenopus* spinal neurons (*111*). Young spinal axons in vitro, like precrossing commissural axons in vivo, are attracted by netrin and are unresponsive to Slit. However, when both cues are applied simultaneously, netrin can still stimulate axon growth but not turning. This is not just a simple matter of repulsion canceling out attraction, because these axons are not repelled by Slit at all, and other attractive responses are not affected. Thus, Slit specifically silences attraction by netrin. This silencing effect is mediated by a direct interaction between the cytoplasmic domains of the Robo and DCC receptors (*111*).

This could explain how attraction by netrin is shut down at the midline, but what turns Slit repulsion on? In flies, Robo receptors are expressed at high levels on commissural axons only after crossing, even though *robo* mRNA is expressed early on (*14*). Robo protein is also synthesized before crossing, but an intracellular sorting receptor (Comm) apparently prevents it from being delivered to the growth cone, targeting the newly synthesized Robo instead for lysosomal degradation (*112*, *113*). Once a commissural axon has crossed the midline, Comm appears to be inactivated, possibly by both transcriptional and posttranscriptional mechanisms. This allows Robo to be delivered to the growth cone, thereby conferring sensitivity to Slit. Thus, it is not the local synthesis or activity of the Robo receptor that is regulated in *Drosophila* commissural axons, but rather its intracellular trafficking. This mechanism may also apply to Robo2 and Robo3. These two Slit receptors are also down-regulated during midline crossing, but must be up-regulated after crossing for axons to select their appropriate pathways on the contralateral side (*24*, *25*).

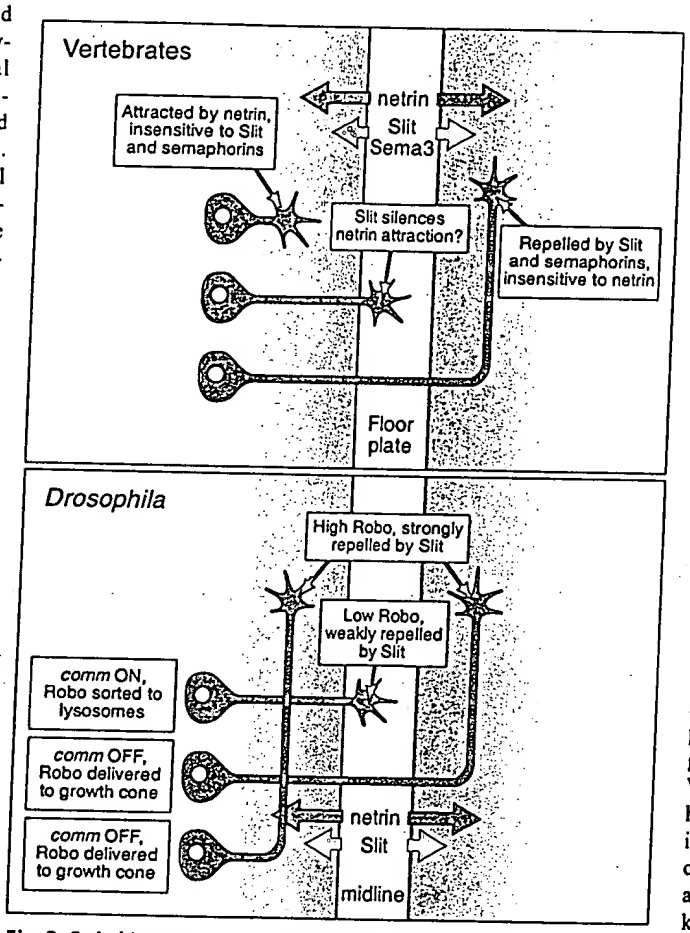


Fig. 3. Switching sensitivity at the midline. As they cross the floor plate, vertebrate commissural axons lose sensitivity to the midline attractant, netrin, and acquire sensitivity to Slit and semaphorin repellents. This switch may be mediated in part by silencing of netrin attraction by Slit. *Drosophila* commissural axons also become sensitive to Slit only after crossing. This appears to reflect Comm's role in regulating the intracellular trafficking of Robo.

Concluding Remarks

Netrins, Slits, semaphorins, and ephrins are not the only guidance cues we know of, and many more undoubtedly still await discovery. Nevertheless, members of these four families have turned up repeatedly in various genetic and biochemical assays and have been found to control a wide range of guidance decisions in vivo. How can so few molecules contribute so much to the correct wiring of the nervous system? Two related principles emerging from these studies seem to be important. First, guidance cues are multifunctional. A single cue can either attract or repel axons, at short or long range, and may even elicit other responses such as branching or an altered sensitivity to other cues. Second, growth cone responses are remarkably plastic, subject to modulation by both instrinsic and extrinsic factors. Together, these mechanisms may underlie much of the diversity in growth cone behavior.

What are the major challenges that still lie ahead? One will be to identify more guidance factors, in particular those that may have more specialized functions, and to figure out how they work. Another challenge will be to gain a better picture of how guidance cues steer growth cones. We now have a few tantalizing glimpses, but are still a long way from a coherent view of growth cone turning. Also, having learned that the outcome of a particular signaling event is essentially unpredictable, the need is now greater than ever to push ahead with the analysis of guidance mechanisms in vivo. We need to know, for example, how the distributions of the various guidance molecules are controlled in space and time, and how each growth cone knows when and how to respond to these cues. The ultimate challenge, after all, is to find out how a comparatively small number of guidance molecules generate such astonishingly complex patterns of neuronal wiring.

References and Notes

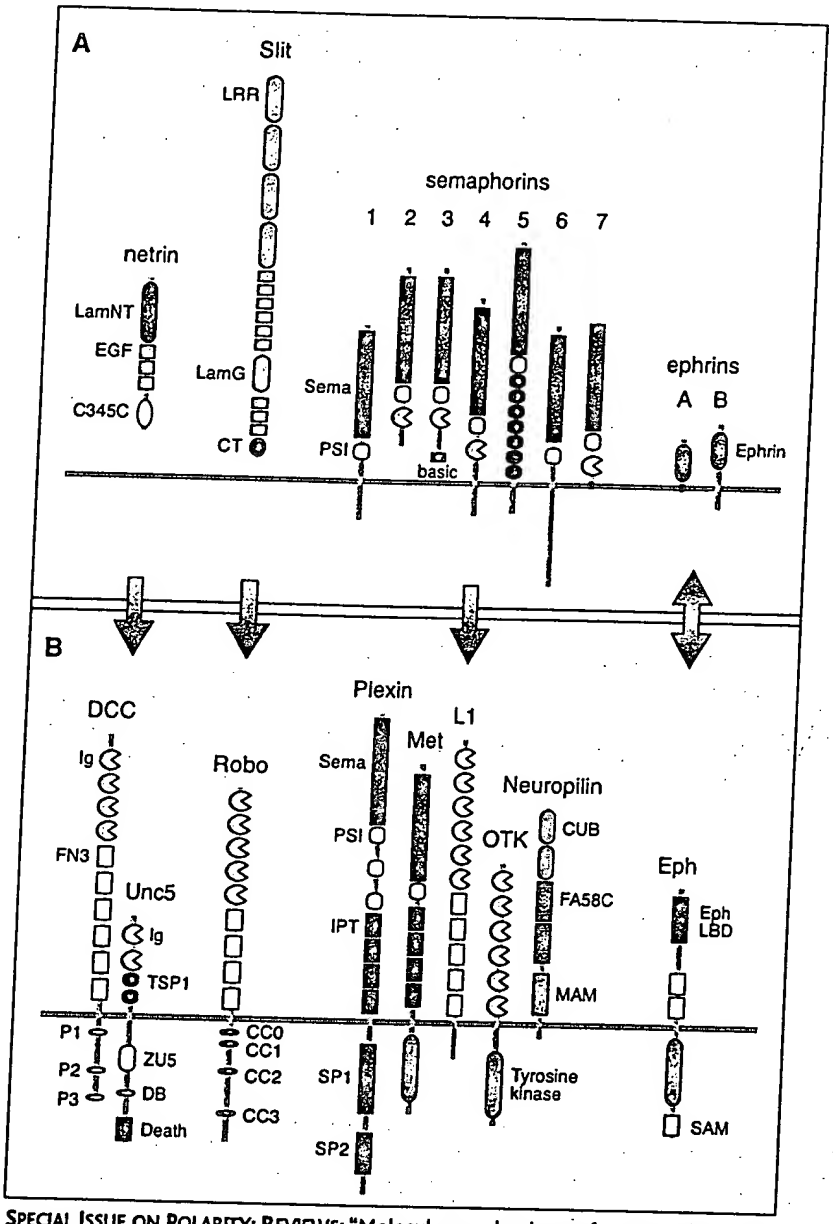
1. M. Tessier-Lavigne, C. S. Goodman, *Science* 274, 1123 (1996).
2. C. S. Goodman, C. J. Shatz, *Cell* 72 (suppl.), 77 (1993).
3. T. E. Kennedy, T. Serafini, J. R. de la Torre, M. Tessier-Lavigne, *Cell* 78, 425 (1994).
4. T. Serafini *et al.* *Cell* 78, 409 (1994).
5. E. M. Hedgecock, J. G. Culotti, D. H. Hall, *Neuron* 4, 61 (1990).
6. N. Ishii, W. G. Wadsworth, B. D. Stern, J. G. Culotti, E. M. Hedgecock, *Neuron* 9, 873 (1992).
7. J. G. Culotti, D. C. Merz, *Curr. Opin. Cell Biol.* 10, 609 (1998).
8. S. A. Colamarino, M. Tessier-Lavigne, *Cell* 81, 621 (1995).
9. M. L. Winberg, K. J. Mitchell, C. S. Goodman, *Cell* 93, 581 (1998).
10. K. Keleman, B. J. Dickson, *Neuron* 32, 605 (2001).

11. K. Hong *et al.*, *Cell* **97**, 927 (1999).
12. K. T. Yee, H. H. Simon, M. Tessier-Lavigne, D. M. O'Leary, *Neuron* **24**, 607 (1999).
13. M. Seeger, G. Tear, D. Ferres-Marco, C. S. Goodman, *Neuron* **10**, 409 (1993).
14. T. Kidd *et al.*, *Cell* **92**, 205 (1998).
15. R. Battye, A. Stevens, J. R. Jacobs, *Development* **126**, 2475 (1999).
16. T. Kidd, K. S. Bland, C. S. Goodman, *Cell* **96**, 785 (1999).
17. K. Brose *et al.*, *Cell* **96**, 795 (1999).
18. H. S. Li *et al.*, *Cell* **96**, 807 (1999).
19. K. H. Wang *et al.*, *Cell* **96**, 771 (1999).
20. J. A. Zallen, B. A. Yi, C. I. Bargmann, *Cell* **92**, 217 (1998).
21. C. Fricke, J. S. Lee, S. Geiger-Rudolph, F. Bonhoeffer, C. B. Chien, *Science* **292**, 507 (2001).
22. H. Hu, *Neuron* **23**, 703 (1999).
23. S. P. Niclou, L. Jia, J. A. Raper, *J. Neurosci.* **20**, 4962 (2000).
24. S. Rajagopalan, V. Vivancos, E. Nicolas, B. J. Dickson, *Cell* **103**, 1033 (2000).
25. J. H. Simpson, K. S. Bland, R. D. Fetter, C. S. Goodman, *Cell* **103**, 1019 (2000).
26. Y. Zou, E. Stoeckli, H. Chen, M. Tessier-Lavigne, *Cell* **102**, 363 (2000).
27. A. S. Plump *et al.*, *Neuron* **33**, 219 (2002).
28. L. D. Hutson, C. B. Chien, *Neuron* **33**, 205 (2002).
29. The sorting out of ipsilateral and contralateral axons at the vertebrate optic chiasm appears to be mediated by ephrin-B ligands (*114*).
30. L. Erskine *et al.*, *J. Neurosci.* **20**, 4975 (2000).
31. T. Ringstedt *et al.*, *J. Neurosci.* **20**, 4983 (2000).
32. A. L. Kolodkin *et al.*, *Neuron* **9**, 831 (1992).
33. Y. Luo, D. Raible, J. A. Raper, *Cell* **75**, 217 (1993).
34. J. A. Raper, *Curr. Opin. Neurobiol.* **10**, 88 (2000).
35. L. Tamagnone *et al.*, *Cell* **99**, 71 (1999).
36. M. L. Winberg *et al.*, *Cell* **95**, 903 (1998).
37. H. J. Cheng *et al.*, *Neuron* **32**, 249 (2001).
38. M. R. Comeau *et al.*, *Immunity* **8**, 473 (1998).
39. J. M. Swiercz, R. Kuner, J. Behrens, S. Offermanns, *Neuron* **35**, 51 (2002).
40. V. Castellani, A. Chedotal, M. Schachner, C. Faivre-Sarrailh, G. Rougon, *Neuron* **27**, 237 (2000).
41. S. Giordano *et al.*, *Nature Cell Biol.* **4**, 720 (2002).
42. M. L. Winberg *et al.*, *Neuron* **32**, 53 (2001).
43. J. T. Wong, S. T. Wong, T. P. O'Connor, *Nature Neurosci.* **2**, 798 (1999).
44. P. J. Roy, H. Zheng, C. E. Warren, J. G. Culotti, *Development* **127**, 755 (2000).
45. T. Fujii *et al.*, *Development* **129**, 2053 (2002).
46. V. E. Ginzburg, P. J. Roy, J. G. Culotti, *Development* **129**, 2065 (2002).
47. R. W. Sperry, *Proc. Natl. Acad. Sci. U.S.A.* **50**, 703 (1963).
48. H. J. Cheng, M. Nakamoto, A. D. Bergemann, J. G. Flanagan, *Cell* **82**, 371 (1995).
49. U. Drescher *et al.*, *Cell* **82**, 359 (1995).
50. D. G. Wilkinson, *Nature Rev. Neurosci.* **2**, 155 (2001).
51. D. A. Feldheim *et al.*, *Neuron* **25**, 563 (2000).
52. A. Brown *et al.*, *Cell* **102**, 77 (2000).
53. R. Hindges, T. McLaughlin, N. Genoud, M. Henkemeyer, D. D. O'Leary, *Neuron* **35**, 475 (2002).
54. F. Mann, S. Ray, W. Harris, C. Holt, *Neuron* **35**, 461 (2002).
55. M. Henkemeyer *et al.*, *Cell* **86**, 35 (1996).
56. K. Kullander *et al.*, *Neuron* **29**, 73 (2001).
57. A. Davy *et al.*, *Genes Dev.* **13**, 3125 (1999).
58. B. Knoll, K. Zarbalis, W. Wurst, U. Drescher, *Development* **128**, 895 (2001).
59. Loss-of-function mutations in the single *Drosophila Eph* gene (*Dek*) do not result in major defects in embryonic CNS axon pathways (*115*), contradicting conclusions drawn from RNA interference studies (*116*).
60. S. E. George, K. Simokat, J. Hardin, A. D. Chisholm, *Cell* **92**, 633 (1998).
61. I. D. Chin-Sang *et al.*, *Cell* **99**, 781 (1999).
62. X. Wang *et al.*, *Mol. Cell* **4**, 903 (1999).
63. J. A. Zallen, S. A. Kirch, C. I. Bargmann, *Development* **126**, 3679 (1999).
64. A. K. Lewis, P. C. Bridgman, *J. Cell Biol.* **119**, 1219 (1992).
65. S. Okabe, N. Hirokawa, *J. Neurosci.* **11**, 1918 (1991).
66. C. H. Lin, E. M. Espreafico, M. S. Mooseker, P. Forscher, *Neuron* **16**, 769 (1996).
67. A. Mallavarapu, T. Mitchison, *J. Cell Biol.* **146**, 1097 (1999).
68. R. W. Gundersen, J. N. Barrett, *J. Cell Biol.* **87**, 546 (1980).
69. J. Q. Zheng, J. J. Wan, M. M. Poo, *J. Neurosci.* **16**, 1140 (1996).
70. C. M. Isbister, T. P. O'Connor, *J. Neurobiol.* **44**, 271 (2000).
71. D. Bentley, A. Toroian-Raymond, *Nature* **323**, 712 (1986).
72. C. B. Chien, D. E. Rosenthal, W. A. Harris, C. E. Holt, *Neuron* **11**, 237 (1993).
73. P. C. Letourneau, *Dev. Biol.* **44**, 92 (1975).
74. S. R. Heidemann, P. Lamoureux, R. E. Buxbaum, *J. Cell Biol.* **111**, 1949 (1990).
75. R. W. Davenport, P. Dou, V. Rehder, S. B. Kater, *Nature* **361**, 721 (1993).
76. E. Tanaka, J. Sabry, *Cell* **83**, 171 (1995).
77. A. W. Schaefer, N. Kabir, P. Forscher, *J. Cell Biol.* **158**, 139 (2002).
78. T. P. O'Connor, J. S. Duerr, D. Bentley, *J. Neurosci.* **10**, 3935 (1990).
79. P. Z. Myers, M. J. Bastiani, *J. Neurosci.* **13**, 127 (1993).
80. M. J. Murray, D. J. Merritt, A. H. Brand, P. M. Whitington, *J. Neurobiol.* **37**, 607 (1998).
81. L. Luo, *Nature Rev. Neurosci.* **1**, 173 (2000).
82. E. A. Lundquist, P. W. Reddien, E. Hartwieg, H. R. Horvitz, C. I. Bargmann, *Development* **128**, 4475 (2001).
83. S. Hakeda-Suzuki *et al.*, *Nature* **416**, 438 (2002).
84. R. S. Kishore, M. V. Sundaram, *Dev. Biol.* **241**, 339 (2002).
85. J. Ng *et al.*, *Nature* **416**, 442 (2002).
86. S. M. Shamah *et al.*, *Cell* **105**, 233 (2001).
87. K. Wong *et al.*, *Cell* **107**, 209 (2001).
88. H. G. Vikis, W. Li, Z. He, K. L. Guan, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 12457 (2000).
89. B. Rohm, B. Rahim, B. Kleiber, I. Hovatta, A. W. Puschel, *FEBS Lett.* **486**, 68 (2000).
90. H. Hing, J. Xiao, N. Harden, L. Lim, S. L. Zipursky, *Cell* **97**, 853 (1999).
91. H. Bito *et al.*, *Neuron* **26**, 431 (2000).
92. A. Colavita, J. G. Culotti, *Dev. Biol.* **194**, 72 (1998).
93. Z. Wills, J. Bateman, C. A. Korey, A. Comer, D. Van Vactor, *Neuron* **22**, 301 (1999).
94. G. J. Bashaw, T. Kidd, D. Murray, T. Pawson, C. S. Goodman, *Cell* **101**, 703 (2000).
95. J. E. Bear *et al.*, *Cell* **109**, 509 (2002).
96. J. E. Bear *et al.*, *Cell* **101**, 717 (2000).
97. L. M. Lanier *et al.*, *Neuron* **22**, 313 (1999).
98. K. Hong, M. Nishiyama, J. Henley, M. Tessier-Lavigne, M. Poo, *Nature* **403**, 93 (2000).
99. J. Q. Zheng, *Nature* **403**, 89 (2000).
100. T. M. Gomez, N. C. Spitzer, *Nature* **397**, 350 (1999).
101. T. M. Gomez, E. Robles, M. Poo, N. C. Spitzer, *Science* **291**, 1983 (2001).
102. G. L. Ming *et al.*, *Neuron* **19**, 1225 (1997).
103. H. J. Song, G. L. Ming, M. M. Poo, *Nature* **388**, 275 (1997).
104. H. Song *et al.*, *Science* **281**, 1515 (1998).
105. D. Shewan, A. Dwivedy, R. Anderson, C. E. Holt, *Nature Neurosci.* **5**, 955 (2002).
106. F. Polleux, T. Morrow, A. Ghosh, *Nature* **404**, 567 (2000).
107. D. Campbell, C. Holt, *Neuron* **32**, 1013 (2001).
108. G. L. Ming *et al.*, *Nature* **417**, 411 (2002).
109. P. A. Brittis, Q. Lu, J. G. Flanagan, *Cell* **110**, 223 (2002).
110. R. Shirasaki, R. Katsumata, F. Murakami, *Science* **279**, 105 (1998).
111. E. Stein, M. Tessier-Lavigne, *Science* **291**, 1928 (2001).
112. K. Keleman *et al.*, *Cell* **110**, 415 (2002).
113. A. Myat *et al.*, *Neuron* **35**, 447 (2002).
114. S. Nakagawa *et al.*, *Neuron* **25**, 599 (2000).
115. M. Boyle, J. B. Thomas, personal communication.
116. T. Bossing, A. H. Brand, *Development* **129**, 18 (2002).

117. I thank M. Tessier-Lavigne, L. Luo, A. Kolodkin, and K. Nasmyth for thoughtful comments on the manuscript; J. Thomas for permission to cite unpublished data; and the many colleagues who have shared their thoughts on these topics. Regrettably, space was too limited to cover more than a few selected areas and to cite all relevant original articles.

ERRATUM

post date 24 January 2003



SPECIAL ISSUE ON POLARITY: REVIEWS: "Molecular mechanisms of axon guidance" by B. J. Dickson (6 Dec. 2002, p. 1959). There were several labeling errors in Fig. 1A. The correct version of the figure appears here.

The Sequence of the Human Genome

J. Craig Venter,[1*] Mark D. Adams,[1] Eugene W. Myers,[1] Peter W. Li,[1] Richard J. Mural,[1] Granger G. Sutton,[1] Hamilton O. Smith,[1] Mark Yandell,[1] Cheryl A. Evans,[1] Robert A. Holt,[1] Jeannine D. Gocayne,[1] Peter Amanatides,[1] Richard M. Ballew,[1] Daniel H. Huson,[1] Jennifer Russo Wortman,[1] Qing Zhang,[1] Chinnappa D. Kodira,[1] Xiangqun H. Zheng,[1] Lin Chen,[1] Marian Skupski,[1] Gangadharan Subramanian,[1] Paul D. Thomas,[1] Jinghui Zhang,[1] George L. Gabor Miklos,[2] Catherine Nelson,[3] Samuel Broder,[1] Andrew G. Clark,[4] Joe Nadeau,[5] Victor A. McKusick,[6] Norton Zinder,[7] Arnold J. Levine,[7] Richard J. Roberts,[8] Mel Simon,[9] Carolyn Slayman,[10] Michael Hunkapiller,[11] Randall Bolanos,[1] Arthur Delcher,[1] Ian Dew,[1] Daniel Fasulo,[1] Michael Flanigan,[1] Liliana Florea,[1] Aaron Halpern,[1] Sridhar Hannenhalli,[1] Saul Kravitz,[1] Samuel Levy,[1] Clark Mobarry,[1] Knut Reinert,[1] Karin Remington,[1] Jane Abu-Threideh,[1] Ellen Beasley,[1] Kendra Biddick,[1] Vivien Bonazzi,[1] Rhonda Brandon,[1] Michele Cargill,[1] Ishwar Chandramouliswaran,[1] Rosane Charlab,[1] Kabir Chaturvedi,[1] Zuoming Deng,[1] Valentina Di Francesco,[1] Patrick Dunn,[1] Karen Eilbeck,[1] Carlos Evangelista,[1] Andrei E. Gabrielian,[1] Weiniu Gan,[1] Wangmao Ge,[1] Fangcheng Gong,[1] Zhiping Gu,[1] Ping Guan,[1] Thomas J. Heiman,[1] Maureen E. Higgins,[1] Rui-Ru Ji,[1] Zhaoxi Ke,[1] Karen A. Ketchum,[1] Zhongwu Lai,[1] Yiding Lei,[1] Zhenya Li,[1] Jiayin Li,[1] Yong Liang,[1] Xiaoying Lin,[1] Fu Lu,[1] Gennady V. Merkulov,[1] Natalia Milshina,[1] Helen M. Moore,[1] Ashwinikumar K Naik,[1] Vaibhav A. Narayan,[1] Beena Neelam,[1] Deborah Nusskern,[1] Douglas B. Rusch,[1] Steven Salzberg,[12] Wei Shao,[1] Bixiong Shue,[1] Jingtao Sun,[1] Zhen Yuan Wang,[1] Aihui Wang,[1] Xin Wang,[1] Jian Wang,[1] Ming-Hui Wei,[1] Ron Wides,[13] Chunlin Xiao,[1] Chunhua Yan,[1] Alison Yao,[1] Jane Ye,[1] Ming Zhan,[1] Weiqing Zhang,[1] Hongyu Zhang,[1] Qi Zhao,[1] Liansheng Zheng,[1] Fei Zhong,[1] Wenyan Zhong,[1] Shiaoping C. Zhu,[1] Shaying Zhao,[12] Dennis Gilbert,[1] Suzanna Baumhueter,[1] Gene Spier,[1] Christine Carter,[1] Anibal Cravchik,[1] Trevor Woodage,[1] Feroze Ali,[1] Huijin An,[1] Aderonke Awe,[1] Danita Baldwin,[1] Holly Baden,[1] Mary Barnstead,[1] Ian Barrow,[1] Karen Beeson,[1] Dana Busam,[1] Amy Carver,[1] Angela Center,[1] Ming Lai Cheng,[1] Liz Curry,[1] Steve Danaher,[1] Lionel Davenport,[1] Raymond Desilets,[1] Susanne Dietz,[1] Kristina Dodson,[1] Lisa Doup,[1] Steven Ferriera,[1] Neha Garg,[1] Andres Gluecksmann,[1] Brit Hart,[1] Jason Haynes,[1] Charles Haynes,[1] Cheryl Heiner,[1] Suzanne Hladun,[1] Damon Hostin,[1] Jarrett Houck,[1] Timothy Howland,[1] Chinyere Ibegwam,[1] Jeffery Johnson,[1] Francis Kalush,[1] Lesley Kline,[1] Shashi Koduru,[1] Amy Love,[1] Felecia Mann,[1] David May,[1] Steven McCawley,[1] Tina McIntosh,[1] Ivy McMullen,[1] Mee Moy,[1] Linda Moy,[1] Brian Murphy,[1] Keith Nelson,[1] Cynthia Pfannkoch,[1] Eric Pratts,[1] Vinita Puri,[1] Hina Qureshi,[1] Matthew Reardon,[1] Robert Rodriguez,[1] Yu-Hui Rogers,[1] Deanna Romblad,[1] Bob Ruhfel,[1] Richard Scott,[1] Cynthia Sitter,[1] Michelle Smallwood,[1] Erin Stewart,[1] Renee Strong,[1] Ellen Suh,[1] Reginald Thomas,[1] Ni Ni Tint,[1] Sukyee Tse,[1] Claire Vech,[1] Gary Wang,[1] Jeremy Wetter,[1] Sherita Williams,[1] Monica Williams,[1] Sandra Windsor,[1] Emily Winn-Deen,[1] Keriellen Wolfe,[1] Jayshree Zaveri,[1] Karena Zaveri,[1] Josep F. Abril,[14] Roderic Guigó,[14] Michael J. Campbell,[1] Kimmen V. Sjolander,[1] Brian Karlak,[1] Anish Kejariwal,[1] Huaiyu Mi,[1] Betty Lazareva,[1] Thomas Hatton,[1] Apurva Narechania,[1] Karen Diemer,[1] Anushya Muruganujan,[1] Nan Guo,[1] Shinji Sato,[1] Vineet Bafna,[1] Sorin Istrail,[1] Ross Lippert,[1] Russell Schwartz,[1] Brian Walenz,[1] Shibu Yooseph,[1] David Allen,[1] Anand Basu,[1] James Baxendale,[1] Louis Blick,[1] Marcelo Caminha,[1] John Carnes-Stine,[1] Parris Caulk,[1] Yen-Hui Chiang,[1] My Coyne,[1] Carl Dahlke,[1] Anne Deslattes Mays,[1] Maria Dombroski,[1] Michael Donnelly,[1] Dale Ely,[1] Shiva Esparham,[1] Carl Fosler,[1] Harold Gire,[1] Stephen Glanowski,[1] Kenneth Glasser,[1] Anna Glodek,[1] Mark Gorokhov,[1] Ken Graham,[1] Barry Gropman,[1] Michael Harris,[1] Jeremy Heil,[1] Scott Henderson,[1] Jeffrey Hoover,[1] Donald Jennings,[1] Catherine Jordan,[1] James Jordan,[1] John Kasha,[1] Leonid Kagan,[1] Cheryl Kraft,[1] Alexander Levitsky,[1] Mark Lewis,[1] Xiangjun Liu,[1] John Lopez,[1] Daniel Ma,[1] William Majoros,[1] Joe McDaniel,[1] Sean Murphy,[1] Matthew Newman,[1] Trung Nguyen,[1] Ngoc Nguyen,[1] Marc Nodell,[1] Sue Pan,[1] Jim Peck,[1] Marshall Peterson,[1] William Rowe,[1] Robert Sanders,[1] John Scott,[1] Michael Simpson,[1] Thomas Smith,[1] Arlan Sprague,[1] Timothy Stockwell,[1] Russell Turner,[1] Eli Venter,[1] Mei Wang,[1] Meiyuan Wen,[1] David Wu,[1] Mitchell Wu,[1] Ashley Xia,[1] Ali Zandieh,[1] Xiaohong Zhu[1]

A 2.91-billion base pair (bp) consensus sequence of the euchromatic portion of the human genome was generated by the whole-genome shotgun sequencing method. The 14.8-billion bp DNA sequence was generated over 9 months from 27,271,853 high-quality sequence reads (5.11-fold coverage of the genome) from both ends of plasmid clones made from the DNA of five individuals. Two assembly strategies—a whole-genome assembly and a regional chromosome assembly—were used, each combining sequence data from Celera and the publicly funded genome effort. The public data were shredded into 550-bp segments to create a 2.9-fold coverage of those genome regions that had been sequenced, without including biases inherent in the cloning and assembly procedure used by the publicly funded group. This brought the effective coverage in the assemblies to eightfold, reducing the number and size of gaps in the final assembly over what would be obtained with 5.11-fold coverage. The two assembly strategies yielded very similar results that largely agree with independent mapping data. The assemblies effectively cover the euchromatic regions of the human chromosomes. More than 90% of the genome is in scaffold assemblies of 100,000 bp or more, and 25% of the genome is in scaffolds of 10 million bp or larger. Analysis of the genome sequence revealed 26,588 protein-encoding transcripts for which there was strong corroborating evidence and an additional ~12,000 computationally derived genes with mouse matches or other weak supporting evidence. Although gene-dense clusters are obvious, almost half the genes are dispersed in low G+C sequence separated by large tracts of apparently noncoding sequence. Only 1.1% of the genome is spanned by exons, whereas 24% is in introns, with 75% of the genome being intergenic DNA. Duplications of segmental blocks, ranging in size up to chromosomal lengths, are abundant throughout the genome and reveal a complex evolutionary history. Comparative genomic analysis indicates vertebrate expansions of genes associated with neuronal function, with tissue-specific developmental regulation, and with the hemostasis and immune systems. DNA sequence comparisons between the consensus sequence and publicly funded genome data provided locations of 2.1 million single-nucleotide polymorphisms (SNPs). A random pair of human haploid genomes differed at a rate of 1 bp per 1250 on average, but there was marked heterogeneity in the level of polymorphism across the genome. Less than 1% of all SNPs resulted in variation in proteins, but the task of determining which SNPs have functional consequences remains an open challenge.

Decoding of the DNA that constitutes the human genome has been widely anticipated for the contribution it will make toward understanding human evolution, the causation of disease, and the interplay between the environment and heredity in defining the human condition. A project with the goal of determining the complete nucleotide sequence of the human genome was first formally proposed in 1985 (*1*). In subsequent years, the idea met with mixed reactions in the scientific community (*2*). However, in 1990, the Human Genome Project (HGP) was officially initiated in the United States under the direction of the National Institutes of Health and the U.S. Department of Energy with a 15-year, $3 billion plan for completing the genome sequence. In 1998 we announced our intention to build a unique genome-sequencing facility, to determine the sequence of the human genome over a 3-year period. Here we report the penultimate milestone along the path toward that goal, a nearly complete sequence of the euchromatic portion of the human genome. The sequencing was performed by a whole-genome random shotgun method with subsequent assembly of the sequenced segments.

The modern history of DNA sequencing began in 1977, when Sanger reported his method for determining the order of nucleotides of DNA using chain-terminating nucleotide analogs (*3*). In the same year, the first human gene was isolated and sequenced (*4*). In 1986, Hood and co-workers (*5*) described an improvement in the Sanger sequencing method that included attaching fluorescent dyes to the nucleotides, which permitted them to be sequentially read by a computer. The first automated DNA sequencer, developed by Applied Biosystems in California in 1987, was shown to be successful when the sequences of two genes were obtained with this new technology (*6*). From early sequencing of human genomic regions (*7*), it became clear that cDNA sequences (which are reverse-transcribed from RNA) would be essential to annotate and validate gene predictions in the human genome. These studies were the basis in part for the development of the expressed sequence tag (EST) method of gene identification (*8*), which is a random selection, very high throughput sequencing approach to characterize cDNA libraries. The EST method led to the rapid discovery and mapping of human genes (*9*). The increasing numbers of human EST sequences necessitated the development of new computer algorithms to analyze large amounts of sequence data, and in 1993 at The Institute for Genomic Research (TIGR), an algorithm was developed that permitted assembly and analysis of hundreds of thousands of ESTs. This algorithm permitted characterization and annotation of human genes on the basis of 30,000 EST assemblies (*10*).

The complete 49-kbp bacteriophage lambda genome sequence was determined by a shotgun restriction digest method in 1982 (*11*). When considering methods for sequencing the smallpox virus genome in 1991 (*12*), a whole-genome shotgun sequencing method was discussed and subsequently rejected owing to the lack of appropriate software tools for genome assembly. However, in 1994, when a microbial genome-sequencing project was contemplated at TIGR, a whole-genome shotgun sequencing approach was considered possible with the TIGR EST assembly algorithm. In 1995, the 1.8-Mbp *Haemophilus influenzae* genome was completed by a whole-genome shotgun sequencing method (*13*). The experience with several subsequent genome-sequencing efforts established the broad applicability of this approach (*14*, *15*).

A key feature of the sequencing approach used for these megabase-size and larger genomes was the use of paired-end sequences (also called mate pairs), derived from subclone libraries with distinct insert sizes and cloning characteristics. Paired-end sequences are sequences 500 to 600 bp in length from both ends of double-stranded DNA clones of prescribed lengths. The success of using end sequences from long segments (18 to 20 kbp) of DNA cloned into bacteriophage lambda in assembly of the microbial genomes led to the suggestion (*16*) of an approach to simulta-

[1]Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA. [2]GenetixXpress, 78 Pacific Road, Palm Beach, Sydney 2108, Australia. [3]Berkeley *Drosophila* Genome Project, University of California, Berkeley, CA 94720, USA. [4]Department of Biology, Penn State University, 208 Mueller Lab, University Park, PA 16802, USA. [5]Department of Genetics, Case Western Reserve University School of Medicine, BRB-630, 10900 Euclid Avenue, Cleveland, OH 44106, USA. [6]Johns Hopkins University School of Medicine, Johns Hopkins Hospital, 600 North Wolfe Street, Blalock 1007, Baltimore, MD 21287–4922, USA. [7]Rockefeller University, 1230 York Avenue, New York, NY 10021–6399, USA. [8]New England BioLabs, 32 Tozer Road, Beverly, MA 01915, USA. [9]Division of Biology, 147-75, California Institute of Technology, 1200 East California Boulevard, Pasadena, CA 91125, USA. [10]Yale University School of Medicine, 333 Cedar Street, P.O. Box 208000, New Haven, CT 06520–8000, USA. [11]Applied Biosystems, 850 Lincoln Centre Drive, Foster City, CA 94404, USA. [12]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA. [13]Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan, 52900 Israel. [14]Grup de Recerca en Informàtica Mèdica, Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, 08003-Barcelona, Catalonia, Spain.

*To whom correspondence should be addressed. E-mail: humangenome@celera.com

neously map and sequence the human genome by means of end sequences from 150-kbp bacterial artificial chromosomes (BACs) (*17*, *18*). The end sequences spanned by known distances provide long-range continuity across the genome. A modification of the BAC end-sequencing (BES) method was applied successfully to complete chromosome 2 from the *Arabidopsis thaliana* genome (*19*).

In 1997, Weber and Myers (*20*) proposed whole-genome shotgun sequencing of the human genome. Their proposal was not well received (*21*). However, by early 1998, as less than 5% of the genome had been sequenced, it was clear that the rate of progress in human genome sequencing worldwide was very slow (*22*), and the prospects for finishing the genome by the 2005 goal were uncertain.

In early 1998, PE Biosystems (now Applied Biosystems) developed an automated, high-throughput capillary DNA sequencer, subsequently called the ABI PRISM 3700 DNA Analyzer. Discussions between PE Biosystems and TIGR scientists resulted in a plan to undertake the sequencing of the human genome with the 3700 DNA Analyzer and the whole-genome shotgun sequencing techniques developed at TIGR (*23*). Many of the principles of operation of a genome-sequencing facility were established in the TIGR facility (*24*). However, the facility envisioned for Celera would have a capacity roughly 50 times that of TIGR, and thus new developments were required for sample preparation and tracking and for whole-genome assembly. Some argued that the required 150-fold scale-up from the *H. influenzae* genome to the human genome with its complex repeat sequences was not feasible (*25*). The *Drosophila melanogaster* genome was thus chosen as a test case for whole-genome assembly on a large and complex eukaryotic genome. In collaboration with Gerald Rubin and the Berkeley *Drosophila* Genome Project, the nucleotide sequence of the 120-Mbp euchromatic portion of the *Drosophila* genome was determined over a 1-year period (*26*–*28*). The *Drosophila* genome-sequencing effort resulted in two key findings: (i) that the assembly algorithms could generate chromosome assemblies with highly accurate order and orientation with substantially less than 10-fold coverage, and (ii) that undertaking multiple interim assemblies in place of one comprehensive final assembly was not of value.

These findings, together with the dramatic changes in the public genome effort subsequent to the formation of Celera (*29*), led to a modified whole-genome shotgun sequencing approach to the human genome. We initially proposed to do 10-fold sequence coverage of the genome over a 3-year period and to make interim assembled sequence data available quarterly. The modifications included a plan to perform random shotgun sequencing to ~5-fold coverage and to use the unordered and unoriented BAC sequence fragments and subassemblies published in GenBank by the publicly funded genome effort (*30*) to accelerate the project. We also abandoned the quarterly announcements in the absence of interim assemblies to report.

Although this strategy provided a reasonable result very early that was consistent with a whole-genome shotgun assembly with eight-fold coverage, the human genome sequence is not as finished as the *Drosophila* genome was with an effective 13-fold coverage. However, it became clear that even with this reduced coverage strategy, Celera could generate an accurately ordered and oriented scaffold sequence of the human genome in less than 1 year. Human genome sequencing was initiated 8 September 1999 and completed 17 June 2000. The first assembly was completed 25 June 2000, and the assembly reported here was completed 1 October 2000. Here we describe the whole-genome random shotgun sequencing effort applied to the human genome. We developed two different assembly approaches for assembling the ~3 billion bp that make up the 23 pairs of chromosomes of the *Homo sapiens* genome. Any GenBank-derived data were shredded to remove potential bias to the final sequence from chimeric clones, foreign DNA contamination, or misassembled contigs. Insofar as a correctly and accurately assembled genome sequence with faithful order and orientation of contigs is essential for an accurate analysis of the human genetic code, we have devoted a considerable portion of this manuscript to the documentation of the quality of our reconstruction of the genome. We also describe our preliminary analysis of the human genetic code on the basis of computational methods. Figure 1 (see fold-out chart associated with this issue; files for each chromosome can be found in Web fig. 1 on *Science* Online at www.sciencemag.org/cgi/content/full/291/5507/1304/DC1) provides a graphical overview of the genome and the features encoded in it. The detailed manual curation and interpretation of the genome are just beginning.

To aid the reader in locating specific analytical sections, we have divided the paper into seven broad sections. A summary of the major results appears at the beginning of each section.

1 Sources of DNA and Sequencing Methods
2 Genome Assembly Strategy and Characterization
3 Gene Prediction and Annotation
4 Genome Structure
5 Genome Evolution
6 A Genome-Wide Examination of Sequence Variations
7 An Overview of the Predicted Protein-Coding Genes in the Human Genome
8 Conclusions

1 Sources of DNA and Sequencing Methods

Summary. This section discusses the rationale and ethical rules governing donor selection to ensure ethnic and gender diversity along with the methodologies for DNA extraction and library construction. The plasmid library construction is the first critical step in shotgun sequencing. If the DNA libraries are not uniform in size, nonchimeric, and do not randomly represent the genome, then the subsequent steps cannot accurately reconstruct the genome sequence. We used automated high-throughput DNA sequencing and the computational infrastructure to enable efficient tracking of enormous amounts of sequence information (27.3 million sequence reads; 14.9 billion bp of sequence). Sequencing and tracking from both ends of plasmid clones from 2-, 10-, and 50-kbp libraries were essential to the computational reconstruction of the genome. Our evidence indicates that the accurate pairing rate of end sequences was greater than 98%.

Various policies of the United States and the World Medical Association, specifically the Declaration of Helsinki, offer recommendations for conducting experiments with human subjects. We convened an Institutional Review Board (IRB) (*31*) that helped us establish the protocol for obtaining and using human DNA and the informed consent process used to enroll research volunteers for the DNA-sequencing studies reported here. We adopted several steps and procedures to protect the privacy rights and confidentiality of the research subjects (donors). These included a two-stage consent process, a secure random alphanumeric coding system for specimens and records, circumscribed contact with the subjects by researchers, and options for off-site contact of donors. In addition, Celera applied for and received a Certificate of Confidentiality from the Department of Health and Human Services. This Certificate authorized Celera to protect the privacy of the individuals who volunteered to be donors as provided in Section 301(d) of the Public Health Service Act 42 U.S.C. 241(d).

Celera and the IRB believed that the initial version of a completed human genome should be a composite derived from multiple donors of diverse ethnic backgrounds Prospective donors were asked, on a voluntary basis, to self-designate an ethnogeographic category (e.g., African-American, Chinese, Hispanic, Caucasian, etc.). We enrolled 21 donors (*32*).

Three basic items of information from each donor were recorded and linked by confidential code to the donated sample: age, sex, and self-designated ethnogeographic group. From females, ~130 ml of whole, heparinized blood was collected. From males, ~130 ml of whole, heparinized blood was

collected, as well as five specimens of semen, collected over a 6-week period. Permanent lymphoblastoid cell lines were created by Epstein-Barr virus immortalization. DNA from five subjects was selected for genomic DNA sequencing: two males and three females—one African-American, one Asian-Chinese, one Hispanic-Mexican, and two Caucasians (see Web fig. 2 on *Science* Online at www.sciencemag.org/cgi/content/291/5507/1304/DC1). The decision of whose DNA to sequence was based on a complex mix of factors, including the goal of achieving diversity as well as technical issues such as the quality of the DNA libraries and availability of immortalized cell lines.

1.1 Library construction and sequencing

Central to the whole-genome shotgun sequencing process is preparation of high-quality plasmid libraries in a variety of insert sizes so that pairs of sequence reads (mates) are obtained, one read from both ends of each plasmid insert. High-quality libraries have an equal representation of all parts of the genome, a small number of clones without inserts, and no contamination from such sources as the mitochondrial genome and *Escherichia coli* genomic DNA. DNA from each donor was used to construct plasmid libraries in one or more of three size classes: 2 kbp, 10 kbp, and 50 kbp (Table 1) (*33*).

In designing the DNA-sequencing process, we focused on developing a simple system that could be implemented in a robust and reproducible manner and monitored effectively (Fig. 2) (*34*).

Current sequencing protocols are based on the dideoxy sequencing method (*35*), which typically yields only 500 to 750 bp of sequence per reaction. This limitation on read length has made monumental gains in throughput a prerequisite for the analysis of large eukaryotic genomes. We accomplished this at the Celera facility, which occupies about 30,000 square feet of laboratory space and produces sequence data continuously at a rate of 175,000 total reads per day. The DNA-sequencing facility is supported by a high-performance computational facility (*36*).

The process for DNA sequencing was modular by design and automated. Intermodule sample backlogs allowed four principal modules to operate independently: (i) library transformation, plating, and colony picking; (ii) DNA template preparation; (iii) dideoxy sequencing reaction set-up and purification; and (iv) sequence determination with the ABI PRISM 3700 DNA Analyzer. Because the inputs and outputs of each module have been carefully matched and sample backlogs are continuously managed, sequencing has proceeded without a single day's interruption since the initiation of the *Drosophila* project in May 1999. The ABI 3700 is a fully automated capillary array sequencer and as such can be operated with a minimal amount of hands-on time, currently estimated at about 15 min per day. The capillary system also facilitates correct associations of sequencing traces with samples through the elimination of manual sample loading and lane-tracking errors associated with slab gels. About 65 production staff were hired and trained, and were rotated on a regular basis through the four production modules. A central laboratory information management system (LIMS) tracked all sample plates by unique bar code identifiers. The facility was supported by a quality control team that performed raw material and in-process testing and a quality assurance group with responsibilities including document control, validation, and auditing of the facility. Critical to the success of the scale-up was the validation of all software and instrumentation before implementation, and production-scale testing of any process changes.

1.2 Trace processing

An automated trace-processing pipeline has been developed to process each sequence file (*37*). After quality and vector trimming, the average trimmed sequence length was 543 bp, and the sequencing accuracy was exponentially distributed with a mean of 99.5% and with less than 1 in 1000 reads being less than 98% accurate (*26*). Each trimmed sequence was screened for matches to contaminants including sequences of vector alone, *E. coli* genomic DNA, and human mitochondrial DNA. The entire read for any sequence with a significant match to a contaminant was discarded. A total of 713 reads matched *E. coli* genomic DNA and 2114 reads matched the human mitochondrial genome.

1.3 Quality assessment and control

The importance of the base-pair level accuracy of the sequence data increases as the size and repetitive nature of the genome to be sequenced increases. Each sequence read must be placed uniquely in the ge-

Table 1. Celera-generated data input into assembly.

	Individual	Number of reads for different insert libraries				Total number of base pairs
		2 kbp	10 kbp	50 kbp	Total	
No. of sequencing reads	A	0	0	2,767,357	2,767,357	1,502,674,851
	B	11,736,757	7,467,755	66,930	19,271,442	10,464,393,006
	C	853,819	881,290	0	1,735,109	942,164,187
	D	952,523	1,046,815	0	1,999,338	1,085,640,534
	F	0	1,498,607	0	1,498,607	813,743,601
	Total	13,543,099	10,894,467	2,834,287	27,271,853	14,808,616,179
Fold sequence coverage (2.9-Gb genome)	A	0	0	0.52	0.52	
	B	2.20	1.40	0.01	3.61	
	C	0.16	1.17	0	0.32	
	D	0.18	0.20	0	0.37	
	F	0	0.28	0	0.28	
	Total	2.54	2.04	0.53	5.11	
Fold clone coverage	A	0	0	18.39	18.39	
	B	2.96	11.26	0.44	14.67	
	C	0.22	1.33	0	1.54	
	D	0.24	1.58	0	1.82	
	F	0	2.26	0	2.26	
	Total	3.42	16.43	18.84	38.68	
Insert size* (mean)	Average	1,951 bp	10,800 bp	50,715 bp		
Insert size* (SD)	Average	6.10%	8.10%	14.90%		
% Mates†	Average	74.50	80.80	75.60		

*Insert size and SD are calculated from assembly of mates on contigs. †% Mates is based on laboratory tracking of sequencing runs.

nome, and even a modest error rate can reduce the effectiveness of assembly. In addition, maintaining the validity of mate-pair information is absolutely critical for the algorithms described below. Procedural controls were established for maintaining the validity of sequence mate-pairs as sequencing reactions proceeded through the process, including strict rules built into the LIMS. The accuracy of sequence data produced by the Celera process was validated in the course of the *Drosophila* genome project (*26*). By collecting data for the entire human genome in a single facility, we were able to ensure uniform quality standards and the cost advantages associated with automation, an economy of scale, and process consistency.

2 Genome Assembly Strategy and Characterization

Summary. We describe in this section the two approaches that we used to assemble the genome. One method involves the computational combination of all sequence reads with shredded data from GenBank to generate an independent, nonbiased view of the genome. The second approach involves clustering all of the fragments to a region or chromosome on the basis of mapping information. The clustered data were then shredded and subjected to computational assembly. Both approaches provided essentially the same reconstruction of assembled DNA sequence with proper order and orientation. The second method provided slightly greater sequence coverage (fewer gaps) and was the principal sequence used for the analysis phase. In addition, we document the completeness and correctness of this assembly process

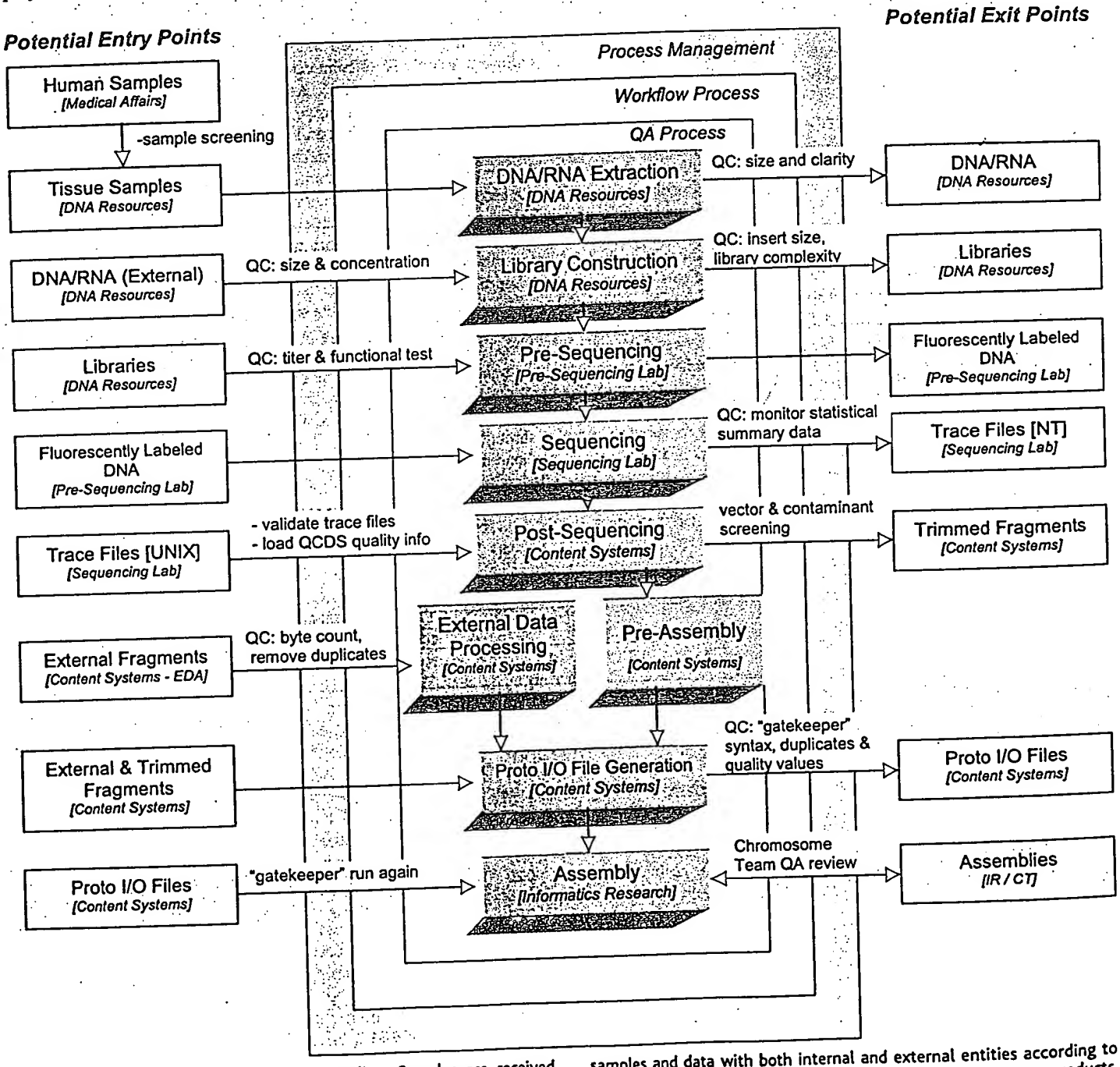


Fig. 2. Flow diagram for sequencing pipeline. Samples are received, selected, and processed in compliance with standard operating procedures, with a focus on quality within and across departments. Each process has defined inputs and outputs with the capability to exchange samples and data with both internal and external entities according to defined quality guidelines. Manufacturing pipeline processes, products, quality control measures, and responsible parties are indicated and are described further in the text.

and provide a comparison to the public genome sequence, which was reconstructed largely by an independent BAC-by-BAC approach. Our assemblies effectively covered the euchromatic regions of the human chromosomes. More than 90% of the genome was in scaffold assemblies of 100,000 bp or greater, and 25% of the genome was in scaffolds of 10 million bp or larger.

Shotgun sequence assembly is a classic example of an inverse problem: given a set of reads randomly sampled from a target sequence, reconstruct the order and the position of those reads in the target. Genome assembly algorithms developed for *Drosophila* have now been extended to assemble the ~25-fold larger human genome. Celera assemblies consist of a set of contigs that are ordered and oriented into scaffolds that are then mapped to chromosomal locations by using known markers. The contigs consist of a collection of overlapping sequence reads that provide a consensus reconstruction for a contiguous interval of the genome. Mate pairs are a central component of the assembly strategy. They are used to produce scaffolds in which the size of gaps between consecutive contigs is known with reasonable precision. This is accomplished by observing that a pair of reads, one of which is in one contig, and the other of which is in another, implies an orientation and distance between the two contigs (Fig. 3). Finally, our assemblies did not incorporate all reads into the final set of reported scaffolds. This set of unincorporated reads is termed "chaff," and typically consisted of reads from within highly repetitive regions, data from other organisms introduced through various routes as found in many genome projects, and data of poor quality or with untrimmed vector.

2.1 Assembly data sets

We used two independent sets of data for our assemblies. The first was a random shotgun data set of 27.27 million reads of average length 543 bp produced at Celera. This consisted largely of mate-pair reads from 16 libraries constructed from DNA samples taken from five different donors. Libraries with insert sizes of 2, 10, and 50 kbp were used. By looking at how mate pairs from a library were positioned in known sequenced stretches of the genome, we were able to characterize the range of insert sizes in each library and determine a mean and standard deviation. Table 1 details the number of reads, sequencing coverage, and clone coverage achieved by the data set. The clone coverage is the coverage of the genome in cloned DNA, considering the entire insert of each clone that has sequence from both ends. The clone coverage provides a measure of the amount of physical DNA coverage of the genome. Assuming a genome size of 2.9 Gbp, the Celera trimmed sequences gave a 5.1× coverage of the genome, and clone coverage was 3.42×, 16.40×, and 18.84× for the 2-, 10-, and 50-kbp libraries, respectively, for a total of 38.7× clone coverage.

The second data set was from the publicly funded Human Genome Project (PFP) and is primarily derived from BAC clones (*30*). The BAC data input to the assemblies came from a download of GenBank on 1 September 2000 (Table 2) totaling 4443.3 Mbp of sequence. The data for each BAC is deposited at one of four levels of completion. Phase 0 data are a set of generally unassembled sequencing reads from a very light shotgun of the BAC, typically less than 1×. Phase 1 data are unordered assemblies of contigs, which we call BAC contigs or bactigs. Phase 2 data are ordered assemblies of bactigs. Phase 3 data are complete BAC sequences. In the past 2 years the PFP has focused on a product of lower quality and completeness, but on a faster time-course, by concentrating on the production of Phase 1 data from a 3× to 4× light-shotgun of each BAC clone.

We screened the bactig sequences for contaminants by using the BLAST algorithm against three data sets: (i) vector sequences in Univec core (*38*), filtered for a 25-bp match at 98% sequence identity at the ends of the sequence and a 30-bp match internal to the sequence; (ii) the nonhuman portion of the High Throughput Genomic (HTG) Seqences division of GenBank (*39*), filtered at 200 bp at 98%; and (iii) the nonredundant nucleotide sequences from GenBank without primate and human virus entries, filtered at 200 bp at 98%. Whenever 25 bp or more of vector was found within 50 bp of the end of a contig, the tip up to the matching vector was excised. Under these criteria we removed 2.6 Mbp of possible contaminant and vector from the Phase 3 data, 61.0 Mbp from the Phase 1 and 2 data, and 16.1 Mbp from the Phase 0 data (Table 2). This left us with a total of 4363.7 Mbp of PFP sequence data 20% finished, 75% rough-draft (Phase 1 and 2), and 5% single sequencing reads (Phase 0). An additional 104,018 BAC end-sequence mate pairs were also downloaded and included in the data sets for both assembly processes (*18*).

2.2 Assembly strategies

Two different approaches to assembly were pursued. The first was a whole-genome assembly process that used Celera data and the PFP data in the form of additional synthetic shotgun data, and the second was a compartmentalized assembly process that first partitioned the Celera and PFP data into sets localized to large chromosomal segments and then performed ab initio shotgun assembly on each set. Figure 4 gives a schematic of the overall process flow.

For the whole-genome assembly, the PFP data was first disassembled or "shredded" into a synthetic shotgun data set of 550-bp reads that form a perfect 2× covering of the bactigs. This resulted in 16.05 million "faux" reads that were sufficient to cover the genome 2.96× because of redundancy in the BAC data set, without incorporating the biases inherent in the PFP assembly process. The combined data set of 43.32 million reads (8×), and all associated mate-pair information, were then subjected to our whole-genome assembly algorithm to produce a reconstruction of the genome. Neither the location of a BAC in the genome nor its assembly of bactigs was used in this process. Bactigs were shredded into reads because we found strong evidence that 2.13% of them were misassembled (*40*). Furthermore, BAC location

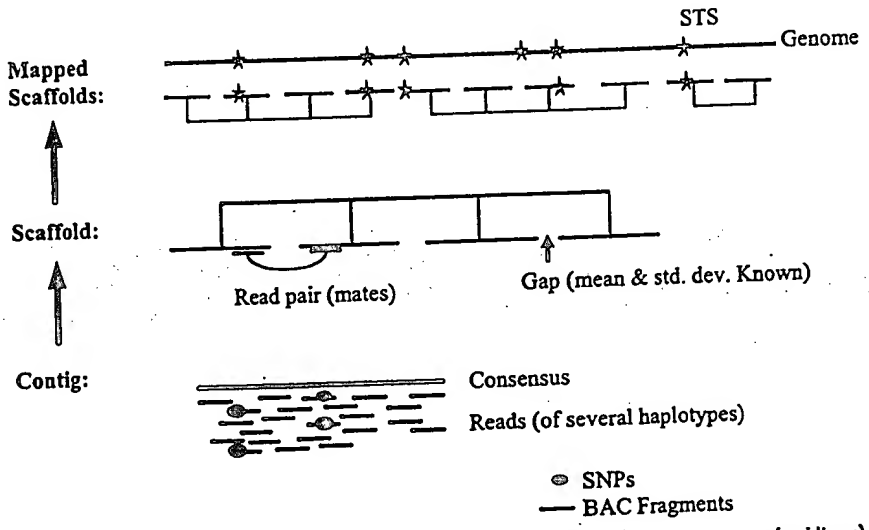


Fig. 3. Anatomy of whole-genome assembly. Overlapping shredded bactig fragments (red lines) and internally derived reads from five different individuals (black lines) are combined to produce a contig and a consensus sequence (green line). Contigs are connected into scaffolds (red) by using mate pair information. Scaffolds are then mapped to the genome (gray line) with STS (blue star) physical map information.

information was ignored because some BACs were not correctly placed on the PFP physical map and because we found strong evidence that at least 2.2% of the BACs contained sequence data that were not part of the given BAC (*41*), possibly as a result of sample-tracking errors (see below). In short, we performed a true, ab initio whole-genome assembly in which we took the expedient of deriving additional sequence coverage, but not mate pairs, assembled bactigs, or genome locality, from some externally generated data.

In the compartmentalized shotgun assembly (CSA), Celera and PFP data were partitioned into the largest possible chromosomal segments or "components" that could be determined with confidence, and then shotgun assembly was applied to each partitioned subset wherein the bactig data were again shredded into faux reads to ensure an independent ab initio assembly of the component. By subsetting the data in this way, the overall computational effort was reduced and the effect of interchromosomal duplications was ameliorated. This also resulted in reconstruction of the genome that was relatively independent of the whole-genome assembly results so that the two assemblies could be compared for consistency. The quality of the partitioning into components was crucial so that different genome regions were not mixed together. We constructed components from (i) the longest scaffolds of the sequence from each BAC and (ii) assembled scaffolds of data unique to Celera's data set. The BAC assemblies were obtained by a combining assembler that used the bactigs and the 5× Celera data mapped to those bactigs as input. This effort was undertaken as an interim step solely because the more accurate and complete the scaffold for a given sequence stretch, the more accurately one can tile the scaffolds into contiguous components on the basis of sequence overlap and mate-pair information. We further visually inspected and curated the scaffold tiling of the components to further increase its accuracy. For the final CSA assembly, all but the partitioning was ignored and an independent, ab initio reconstruction of the sequence in each component was obtained by applying our whole-genome assembly algorithm to the partitioned, relevant Celera data and the shredded, faux reads of the partitioned, relevant bactig data.

Table 2. GenBank data input into assembly.

Center	Statistics	Completion phase sequence		
		0	1 and 2	3
Whitehead Institute/ MIT Center for Genome Research, USA	Number of accession records	2,825	6,533	363
	Number of contigs	243,786	138,023	363
	Total base pairs	194,490,158	1,083,848,245	48,829,358
	Total vector masked (bp)	1,553,597	875,618	2,202
	Total contaminant masked (bp)	13,654,482	4,417,055	98,028
	Average contig length (bp)	798	7,853	134,516
Washington University, USA	Number of accession records	19	3,232	1,300
	Number of contigs	2,127	61,812	1,300
	Total base pairs	1,195,732	561,171,788	164,214,395
	Total vector masked (bp)	21,604	270,942	8,287
	Total contaminant masked (bp)	22,469	1,476,141	469,487
	Average contig length (bp)	562	9,079	126,319
Baylor College of Medicine, USA	Number of accession records	0	1,626	363
	Number of contigs	0	44,861	363
	Total base pairs	0	265,547,066	49,017,104
	Total vector masked (bp)	0	218,769	4,960
	Total contaminant masked (bp)	0	1,784,700	485,137
	Average contig length (bp)	0	5,919	135,033
Production Sequencing Facility, DOE Joint Genome Institute, USA	Number of accession records	135	2,043	754
	Number of contigs	7,052	34,938	754
	Total base pairs	8,680,214	294,249,631	60,975,328
	Total vector masked (bp)	22,644	162,651	7,274
	Total contaminant masked (bp)	665,818	4,642,372	118,387
	Average contig length (bp)	1,231	8,422	80,867
The Institute of Physical and Chemical Research (RIKEN), Japan	Number of accession records	0	1,149	300
	Number of contigs	0	25,772	300
	Total base pairs	0	182,812,275	20,093,926
	Total vector masked (bp)	0	203,792	2,371
	Total contaminant masked (bp)	0	308,426	27,781
	Average contig length (bp)	0	7,093	66,978
Sanger Centre, UK	Number of accession records	0	4,538	2,599
	Number of contigs	0	74,324	2,599
	Total base pairs	0	689,059,692	246,118,000
	Total vector masked (bp)	0	427,326	25,054
	Total contaminant masked (bp)	0	2,066,305	374,561
	Average contig length (bp)	0	9,271	94,697
Others*	Number of accession records	42	1,894	3,458
	Number of contigs	5,978	29,898	3,458
	Total base pairs	5,564,879	283,358,877	246,474,157
	Total vector masked (bp)	57,448	279,477	32,136
	Total contaminant masked (bp)	575,366	1,616,665	1,791,849
	Average contig length (bp)	931	9,478	71,277
All centers combined†	Number of accession records	3,021	21,015	9,137
	Number of contigs	258,943	409,628	9,137
	Total base pairs	209,930,983	3,360,047,574	835,722,268
	Total vector masked (bp)	1,655,293	2,438,575	82,284
	Total contaminant masked (bp)	14,918,135	16,311,664	3,365,230
	Average contig length (bp)	811	8,203	91,466

*Other centers contributing at least 0.1% of the sequence include: Chinese National Human Genome Center; Genomanalyse Gesellschaft fuer Biotechnologische Forschung mbH; Genome Therapeutics Corporation; GENOSCOPE; Chinese Academy of Sciences; Institute of Molecular Biotechnology; Keio University School of Medicine; Lawrence Livermore National Laboratory; Cold Spring Harbor Laboratory; Los Alamos National Laboratory; Max-Planck Institut fuer Molekulare, Genetik; Japan Science and Technology Corporation; Stanford University; The Institute for Genomic Research; The Institute of Physical and Chemical Research, Gene Bank; The University of Oklahoma; University of Texas Southwestern Medical Center, University of Washington. †The 4,405,700,825 bases contributed by all centers were shredded into faux reads resulting in 2.96× coverage of the genome.

2.3 Whole-genome assembly

The algorithms used for whole-genome assembly (WGA) of the human genome were enhancements to those used to produce the sequence of the *Drosophila* genome reported in detail in (*28*).

The WGA assembler consists of a pipeline composed of five principal stages: Screener, Overlapper, Unitigger, Scaffolder, and Repeat Resolver, respectively. The Screener finds and marks all microsatellite repeats with more than a 6-bp element, and screens out known interspersed repeat elements, including Alu, Line, and ribosomal DNA. Marked regions get searched for overlaps, while screened regions do not get searched, but can be part of an overlap that involves unscreened matching segments.

The Overlapper compares every read against every other read in search of complete end-to-end overlaps of at least 40 bp and with no more than 6% differences in the match. Because all data are scrupulously vector-trimmed, the Overlapper can insist on complete overlap matches. Computing the set of all overlaps took roughly 10,000 CPU hours with a suite of four-processor Alpha SMPs with 4 gigabytes of RAM. This took 4 to 5 days in elapsed time with 40 such machines operating in parallel.

Every overlap computed above is statistically a 1-in-10^{17} event and thus not a coincidental event. What makes assembly combinatorially difficult is that while many overlaps are actually sampled from overlapping regions of the genome, and thus imply that the sequence reads should be assembled together, even more overlaps are actually from two distinct copies of a low-copy repeated element not screened above, thus constituting an error if put together. We call the former "true overlaps" and the latter "repeat-induced overlaps." The assembler must avoid choosing repeat-induced overlaps, especially early in the process.

We achieve this objective in the Unitigger. We first find all assemblies of reads that appear to be uncontested with respect to all other reads. We call the contigs formed from these subassemblies unitigs (for uniquely assembled contigs). Formally, these unitigs are the uncontested interval subgraphs of the graph of all overlaps (*42*). Unfortunately, although empirically many of these assemblies are correct (and thus involve only true overlaps), some are in fact collections of reads from several copies of a repetitive element that have been overcollapsed into a single subassembly. However, the overcollapsed unitigs are easily identified because their average coverage depth is too high to be consistent with the overall level of sequence coverage. We developed a simple statistical discriminator that gives the logarithm of the odds ratio that a unitig is composed of unique DNA or of a repeat consisting of two or more copies. The discriminator, set to a sufficiently stringent threshold, identifies a subset of the unitigs that we are certain are correct. In addition, a second, less stringent threshold identifies a subset of remaining unitigs very likely to be correctly assembled, of which we select those that will consistently scaffold (see below), and thus are again almost certain to be correct. We call the union of these two sets U-unitigs. Empirically, we found from a 6× simulated shotgun of human chromosome 22 that we get U-unitigs covering 98% of the stretches of unique DNA that are >2 kbp long. We are further able to identify the boundary of the start of a repetitive element at the ends of a U-unitig and leverage this so that U-unitigs span more than 93% of all singly interspersed Alu elements and other 100-to 400-bp repetitive segments.

The result of running the Unitigger was thus a set of correctly assembled subcontigs covering an estimated 73.6% of the human genome. The Scaffolder then proceeded to use mate-pair information to link these together into scaffolds. When there are two or more mate pairs that imply that a given pair of U-unitigs are at a certain distance and orientation with respect to each other, the probability of this being wrong is again roughly 1 in 10^{10}, assuming that mate pairs are false less than 2% of the time. Thus, one can with high confidence link together all U-unitigs that are linked by at least two 2- or 10-kbp mate pairs producing intermediate-sized scaffolds that are then recursively linked together by confirming 50-kbp mate pairs and BAC end sequences. This process yielded scaffolds that are on the order of megabase pairs in size with gaps between their contigs that generally correspond to repetitive elements and occasionally to small sequencing gaps. These scaffolds reconstruct the majority of the unique sequence within a genome.

For the *Drosophila* assembly, we engaged in a three-stage repeat resolution strategy where each stage was progressively more aggressive and thus more likely to make a mistake. For the human assembly, we continued to use the first "Rocks" substage where all unitigs with a good, but not definitive, discriminator score are placed in a scaffold gap. This was done with the condition that two or more mate pairs with one of their reads already in the scaffold unambiguously place the unitig in the given gap. We estimate the probability of inserting a unitig into an incorrect gap with this strategy to be less than 10^{-7} based on a probabilistic analysis.

We revised the ensuing "Stones" substage of the human assembly, making it more like the mechanism suggested in our earlier work (*43*). For each gap, every read R that is placed in the gap by virtue of its mated pair M being in a contig of the scaffold and implying R's placement is collected. Celera's mate-pairing information is correct more than 99% of the time. Thus, almost every, but not all, of the reads in the set belong in the gap, and when a read does not belong it rarely agrees with the remainder of the reads. Therefore, we simply assemble this set of reads within the gap, eliminating any reads that conflict with the assembly. This operation proved much more reliable than the one it replaced for the *Drosophila* assembly; in the assembly of a simulated shotgun data set of human chromo-

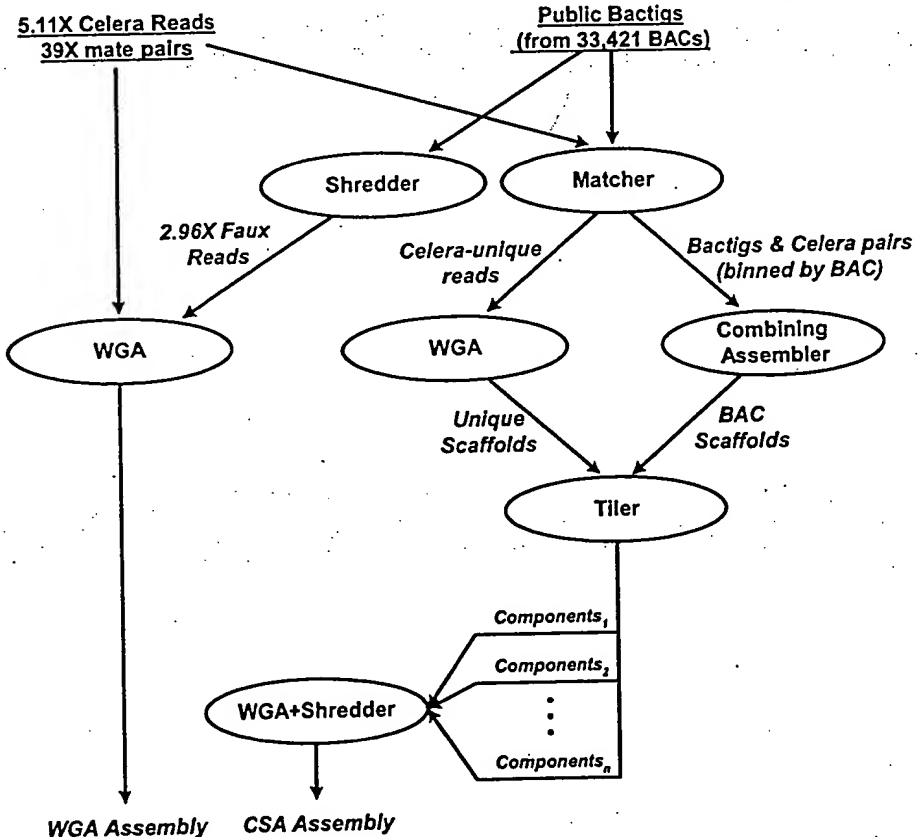


Fig. 4. Architecture of Celera's two-pronged assembly strategy. Each oval denotes a computation process performing the function indicated by its label, with the labels on arcs between ovals describing the nature of the objects produced and/or consumed by a process. This figure summarizes the discussion in the text that defines the terms and phrases used.

some 22, all stones were placed correctly.

The final method of resolving gaps is to fill them with assembled BAC data that cover the gap. We call this external gap "walking." We did not include the very aggressive "Pebbles" substage described in our *Drosophila* work, which made enough mistakes so as to produce repeat reconstructions for long interspersed elements whose quality was only 99.62% correct. We decided that for the human genome it was philosophically better not to introduce a step that was certain to produce less than 99.99% accuracy. The cost was a somewhat larger number of gaps of somewhat larger size.

At the final stage of the assembly process, and also at several intermediate points, a consensus sequence of every contig is produced. Our algorithm is driven by the principle of maximum parsimony, with quality-value–weighted measures for evaluating each base. The net effect is a Bayesian estimate of the correct base to report at each position. Consensus generation uses Celera data whenever it is present. In the event that no Celera data cover a given region, the BAC data sequence is used.

A key element of achieving a WGA of the human genome was to parallelize the Overlapper and the central consensus sequence–constructing subroutines. In addition, memory was a real issue—a straightforward application of the software we had built for *Drosophila* would have required a computer with a 600-gigabyte RAM. By making the Overlapper and Unitigger incremental, we were able to achieve the same computation with a maximum of instantaneous usage of 28 gigabytes of RAM. Moreover, the incremental nature of the first three stages allowed us to continually update the state of this part of the computation as data were delivered and then perform a 7-day run to complete Scaffolding and Repeat Resolution whenever desired. For our assembly operations, the total compute infrastructure consists of 10 four-processor SMPs with 4 gigabytes of memory per cluster (Compaq's ES40, Regatta) and a 16-processor NUMA machine with 64 gigabytes of memory (Compaq's GS160, Wildfire). The total compute for a run of the assembler was roughly 20,000 CPU hours.

The assembly of Celera's data, together with the shredded bactig data, produced a set of scaffolds totaling 2.848 Gbp in span and consisting of 2.586 Gbp of sequence. The chaff, or set of reads not incorporated in the assembly, numbered 11.27 million (26%), which is consistent with our experience for *Drosophila*. More than 84% of the genome was covered by scaffolds >100 kbp long, and these averaged 91% sequence and 9% gaps with a total of 2.297 Gbp of sequence. There were a total of 93,857 gaps among the 1637 scaffolds >100 kbp. The average scaffold size was 1.5 Mbp, the average contig size was 24.06 kbp, and the average gap size was 2.43 kbp, where the distribution of each was essentially exponential More than 50% of all gaps were less than 50(bp long, >62% of all gaps were less than 1 kbp long, and no gap was >100 kbp long. Similarly, more than 65% of the sequence is in contigs >30 kbp, more than 31% is in contigs >100 kbp, and the largest contig was 1.22 Mbp long. Table 3 gives detailed summary statistics for the structure of this assembly with a direct comparison to the compartmentalized shotgun assembly.

2.4 Compartmentalized shotgun assembly

In addition to the WGA approach, we pursued a localized assembly approach that was intended to subdivide the genome into segments, each of which could be shotgun assembled individually. We expected that this would help in resolution of large interchromosomal duplications and improve the statistics for calculating U-unitigs. The compartmentalized assembly process involved clustering Celera reads and bactigs into large, multiple megabase regions of the genome, and then running the WGA assembler on the Celera data and shredded, faux reads obtained from the bactig data.

The first phase of the CSA strategy was to separate Celera reads into those that matched the BAC contigs for a particular PFP BAC entry, and those that did not match any public data. Such matches must be guaranteed to

Table 3. Scaffold statistics for whole-genome and compartmentalized shotgun assemblies.

	Scaffold size				
	All	>30 kbp	>100 kbp	>500 kbp	>1000 kbp
	Compartmentalized shotgun assembly				
No. of bp in scaffolds (including intrascaffold gaps)	2,905,568,203	2,748,892,430	2,700,489,906	2,489,357,260	2,248,689,128
No. of bp in contigs	2,653,979,733	2,524,251,302	2,491,538,372	2,320,648,201	2,106,521,902
No. of scaffolds	53,591	2,845	1,935	1,060	721
No. of contigs	170,033	112,207	107,199	93,138	82,009
No. of gaps	116,442	109,362	105,264	92,078	81,288
No. of gaps ≤1 kbp	72,091	69,175	67,289	59,915	53,354
Average scaffold size (bp)	54,217	966,219	1,395,602	2,348,450	3,118,848
Average contig size (bp)	15,609	22,496	23,242	24,916	25,686
Average intrascaffold gap size (bp)	2,161	2,054	1,985	1,832	1,749
Largest contig (bp)	1,988,321	1,988,321	1,988,321	1,988,321	1,988,321
% of total contigs	100	95	94	87	79
	Whole-genome assembly				
No. of bp in scaffolds (including intrascaffold gaps)	2,847,890,390	2,574,792,618	2,525,334,447	2,328,535,466	2,140,943,032
No. of bp in contigs	2,586,634,108	2,334,343,339	2,297,678,935	2,143,002,184	1,983,305,432
No. of scaffolds	118,968	2,507	1,637	818	554
No. of contigs	221,036	99,189	95,494	84,641	76,285
No. of gaps	102,068	96,682	93,857	83,823	75,731
No. of gaps ≤1 kbp	62,356	60,343	59,156	54,079	49,592
Average scaffold size (bp)	23,938	1,027,041	1,542,660	2,846,620	3,864,518
Average contig size (bp)	11,702	23,534	24,061	25,319	25,999
Average intrascaffold gap size (bp)	2,560	2,487	2,426	2,213	2,082
Largest contig (bp)	1,224,073	1,224,073	1,224,073	1,224,073	1,224,073
% of total contigs	100	90	89	83	77

properly place a Celera read, so all reads were first masked against a library of common repetitive elements, and only matches of at least 40 bp to unmasked portions of the read constituted a hit. Of Celera's 27.27 million reads, 20.76 million matched a bactig and another 0.62 million reads, which did not have any matches, were nonetheless identified as belonging in the region of the bactig's BAC because their mate matched the bactig. Of the remaining reads, 2.92 million were completely screened out and so could not be matched, but the other 2.97 million reads had unmasked sequence totaling 1.189 Gbp that were not found in the GenBank data set. Because the Celera data are 5.11× redundant, we estimate that 240 Mbp of unique Celera sequence is not in the GenBank data set.

In the next step of the CSA process, a combining assembler took the relevant 5× Celera reads and bactigs for a BAC entry, and produced an assembly of the combined data for that locale. These high-quality sequence reconstructions were a transient result whose utility was simply to provide more reliable information for the purposes of their tiling into sets of overlapping and adjacent scaffold sequences in the next step. In outline, the combining assembler first examines the set of matching Celera reads to determine if there are excessive pileups indicative of unscreened repetitive elements. Wherever these occur, reads in the repeat region whose mates have not been mapped to consistent positions are removed. Then all sets of mate pairs that consistently imply the same relative position of two bactigs are bundled into a link and weighted according to the number of mates in the bundle. A "greedy" strategy then attempts to order the bactigs by selecting bundles of mate-pairs in order of their weight. A selected mate-pair bundle can tie together two formative scaffolds. It is incorporated to form a single scaffold only if it is consistent with the majority of links between contigs of the scaffold. Once scaffolding is complete, gaps are filled by the "Stones" strategy described above for the WGA assembler.

The GenBank data for the Phase 1 and 2 BACs consisted of an average of 19.8 bactigs per BAC of average size 8099 bp. Application of the combining assembler resulted in individual Celera BAC assemblies being put together into an average of 1.83 scaffolds (median of 1 scaffold) consisting of an average of 8.57 contigs of average size 18,973 bp. In addition to defining order and orientation of the sequence fragments, there were 57% fewer gaps in the combined result. For Phase 0 data, the average GenBank entry consisted of 91.52 reads of average length 784 bp. Application of the combining assembler resulted in an average of 54.8 scaffolds consisting of an average of 58.1 contigs of average size 873 bp. Basically, some small amount of assembly took place, but not enough Celera data were matched to truly assemble the 0.5× to 1× data set represented by the typical Phase 0 BACs. The combining assembler was also applied to the Phase 3 BACs for SNP identification, confirmation of assembly, and localization of the Celera reads. The phase 0 data suggest that a combined whole-genome shotgun data set and 1× light-shotgun of BACs will not yield good assembly of BAC regions; at least 3× light-shotgun of each BAC is needed.

The 5.89 million Celera fragments not matching the GenBank data were assembled with our whole-genome assembler. The assembly resulted in a set of scaffolds totaling 442 Mbp in span and consisting of 326 Mbp of sequence. More than 20% of the scaffolds were >5 kbp long, and these averaged 63% sequence and 27% gaps with a total of 302 Mbp of sequence. All scaffolds >5 kbp were forwarded along with all scaffolds produced by the combining assembler to the subsequent tiling phase.

At this stage, we typically had one or two scaffolds for every BAC region constituting at least 95% of the relevant sequence, and a collection of disjoint Celera-unique scaffolds. The next step in developing the genome components was to determine the order and overlap tiling of these BAC and Celera-unique scaffolds across the genome. For this, we used Celera's 50-kbp mate-pairs information, and BAC-end pairs (*18*) and sequence tagged site (STS) markers (*44*) to provide long-range guidance and chromosome separation. Given the relatively manageable number of scaffolds, we chose not to produce this tiling in a fully automated manner, but to compute an initial tiling with a good heuristic and then use human curators to resolve discrepancies or missed join opportunities. To this end, we developed a graphical user interface that displayed the graph of tiling overlaps and the evidence for each. A human curator could then explore the implication of mapped STS data, dot-plots of sequence overlap, and a visual display of the mate-pair evidence supporting a given choice. The result of this process was a collection of "components," where each component was a tiled set of BAC and Celera-unique scaffolds that had been curator-approved. The process resulted in 3845 components with an estimated span of 2.922 Gbp.

In order to generate the final CSA, we assembled each component with the WGA algorithm. As was done in the WGA process, the bactig data were shredded into a synthetic 2× shotgun data set in order to give the assembler the freedom to independently assemble the data. By using faux reads rather than bactigs, the assembly algorithm could correct errors in the assembly of bactigs and remove chimeric content in a PFP data entry. Chimeric or contaminating sequence (from another part of the genome) would not be incorporated into the reassembly of the component because it did not belong there. In effect, the previous steps in the CSA process served only to bring together Celera fragments and PFP data relevant to a large contiguous segment of the genome, wherein we applied the assembler used for WGA to produce an ab initio assembly of the region.

WGA assembly of the components resulted in a set of scaffolds totaling 2.906 Gbp in span and consisting of 2.654 Gbp of sequence. The chaff, or set of reads not incorporated into the assembly, numbered 6.17 million, or 22%. More than 90.0% of the genome was covered by scaffolds spanning >100 kbp long, and these averaged 92.2% sequence and 7.8% gaps with a total of 2.492 Gbp of sequence. There were a total of 105,264 gaps among the 107,199 contigs that belong to the 1940 scaffolds spanning >100 kbp. The average scaffold size was 1.4 Mbp, the average contig size was 23.24 kbp, and the average gap size was 2.0 kbp where each distribution of sizes was exponential. As such, averages tend to be underrepresentative of the majority of the data. Figure 5 shows a histogram of the bases in scaffolds of various size ranges. Consider also that more than 49% of all gaps were <500 bp long, more than 62% of all gaps were <1 kbp, and all gaps are <100 kbp long. Similarly, more than 73% of the sequence is in contigs > 30 kbp, more than 49% is in contigs >100 kbp, and the largest contig was 1.99 Mbp long. Table 3 provides summary statistics for the structure of this assembly with a direct comparison to the WGA assembly.

2.5 Comparison of the WGA and CSA scaffolds

Having obtained two assemblies of the human genome via independent computational processes (WGA and CSA), we compared scaffolds from the two assemblies as another means of investigating their completeness, consistency, and contiguity. From each assembly, a set of reference scaffolds containing at least 1000 fragments (Celera sequencing reads or bactig shreds) was obtained; this amounted to 2218 WGA scaffolds and 1717 CSA scaffolds, for a total of 2.087 Gbp and 2.474 Gbp. The sequence of each reference scaffold was compared to the sequence of all scaffolds from the other assembly with which it shared at least 20 fragments or at least 20% of the fragments of the smaller scaffold. For each such comparison, all matches of at least 200 bp with at most 2% mismatch were tabulated.

From this tabulation, we estimated the amount of unique sequence in each assembly in two ways. The first was to determine the number of bases of each assembly that were

not covered by a matching segment in the other assembly. Some 82.5 Mbp of the WGA (3.95%) was not covered by the CSA, whereas 204.5 Mbp (8.26%) of the CSA was not covered by the WGA. This estimate did not require any consistency of the assemblies or any uniqueness of the matching segments. Thus, another analysis was conducted in which matches of less than 1 kbp between a pair of scaffolds were excluded unless they were confirmed by other matches having a consistent order and orientation. This gives some measure of consistent coverage: 1.982 Gbp (95.00%) of the WGA is covered by the CSA, and 2.169 Gbp (87.69%) of the CSA is covered by the WGA by this more stringent measure.

The comparison of WGA to CSA also permitted evaluation of scaffolds for structural inconsistencies. We looked for instances in which a large section of a scaffold from one assembly matched only one scaffold from the other assembly, but failed to match over the full length of the overlap implied by the matching segments. An initial set of candidates was identified automatically, and then each candidate was inspected by hand. From this process, we identified 31 instances in which the assemblies appear to disagree in a nonlocal fashion. These cases are being further evaluated to determine which assembly is in error and why.

In addition, we evaluated local inconsistencies of order or orientation. The following results exclude cases in which one contig in one assembly corresponds to more than one overlapping contig in the other assembly (as long as the order and orientation of the latter agrees with the positions they match in the former). Most of these small rearrangements involved segments on the order of hundreds of base pairs and rarely >1 kbp. We found a total of 295 kbp (0.012%) in the CSA assemblies that were locally inconsistent with the WGA assemblies, whereas 2.108 Mbp (0.11%) in the WGA assembly were inconsistent with the CSA assembly.

The CSA assembly was a few percentage points better in terms of coverage and slightly more consistent than the WGA, because it was in effect performing a few thousand shotgun assemblies of megabase-sized problems, whereas the WGA is performing a shotgun assembly of a gigabase-sized problem. When one considers the increase of two-and-a-half orders of magnitude in problem size, the information loss between the two is remarkably small. Because CSA was logistically easier to deliver and the better of the two results available at the time when downstream analyses needed to be begun, all subsequent analysis was performed on this assembly.

2.6 Mapping scaffolds to the genome

The final step in assembling the genome was to order and orient the scaffolds on the chromosomes. We first grouped scaffolds together on the basis of their order in the components from CSA. These grouped scaffolds were reordered by examining residual mate-pairing data between the scaffolds. We next mapped the scaffold groups onto the chromosome using physical mapping data. This step depends on having reliable high-resolution map information such that each scaffold will overlap multiple markers. There are two genome-wide types of map information available: high-density STS maps and fingerprint maps of BAC clones developed at Washington University (*45*). Among the genome-wide STS maps, GeneMap99 (GM99) has the most markers and therefore was most useful for mapping scaffolds. The two different mapping approaches are complementary to one another. The fingerprint maps should have better local order because they were built by comparison of overlapping BAC clones. On the other hand, GM99 should have a more reliable long-range order, because the framework markers were derived from well-validated genetic maps. Both types of maps were used as a reference for human curation of the components that were the input to the regional assembly, but they did not determine the order of sequences produced by the assembler.

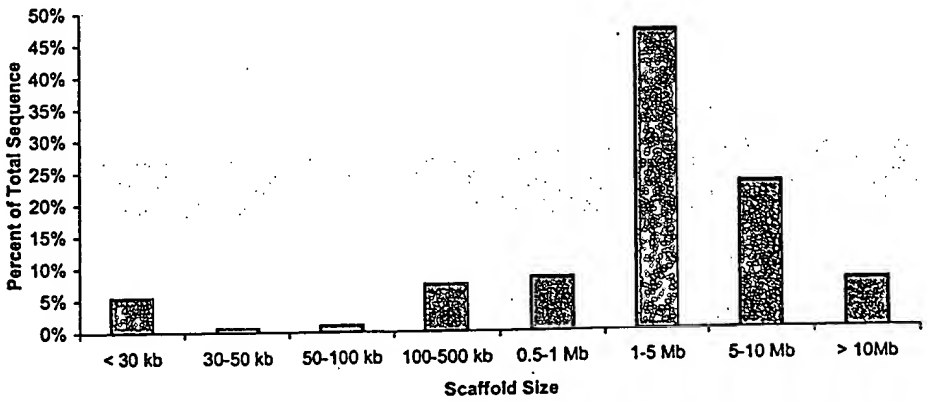


Fig. 5. Distribution of scaffold sizes of the CSA. For each range of scaffold sizes, the percent of total sequence is indicated.

In order to determine the effectiveness of the fingerprint maps and GM99 for mapping scaffolds, we first examined the reliability of these maps by comparison with large scaffolds. Only 1% of the STS markers on the 10 largest scaffolds (those >9 Mbp) were mapped on a different chromosome on GM99. Two percent of the STS markers disagreed in position by more than five framework bins. However, for the fingerprint maps, a 2% chromosome discrepancy was observed, and on average 23.8% of BAC locations in the scaffold sequence disagreed with fingerprint map placement by more than five BACs. When further examining the source of discrepancy, it was found that most of the discrepancy came from 4 of the 10 scaffolds, indicating this there is variation in the quality of either the map or the scaffolds. All four scaffolds were assembled, as well as the other six, as judged by clone coverage analysis, and showed the same low discrepancy rate to GM99, and thus we concluded that the fingerprint map global order in these cases was not reliable. Smaller scaffolds had a higher discordance rate with GM99 (4.21% of STSs were discordant by more than five framework bins), but a lower discordance rate with the fingerprint maps (11% of BACs disagreed with fingerprint maps by more than five BACs). This observation agrees with the clone coverage analysis (*46*) that Celera scaffold construction was better supported by long-range mate pairs in larger scaffolds than in small scaffolds.

We created two orderings of Celera scaffolds on the basis of the markers (BAC or STS) on these maps. Where the order of scaffolds agreed between GM99 and the WashU BAC map, we had a high degree of confidence that that order was correct; these scaffolds were termed "anchor scaffolds." Only scaffolds with a low overall discrepancy rate with both maps were considered anchor scaffolds. Scaffolds in GM99 bins were allowed to permute in their order to match WashU ordering, provided they did not violate their framework orders. Orientation of individual scaffolds was determined by the presence of multiple mapped markers with consistent order. Scaffolds with only one marker have insufficient information to assign orientation. We found 70.1% of the genome in anchored scaffolds, more than 99% of which are also oriented (Table 4). Because GM99 is of lower resolution than the WashU map, a number of scaffolds without STS matches could be ordered relative to the anchored scaffolds because they included sequence from the same or adjacent BACs on the WashU map. On the other hand, because of occasional WashU global ordering discrepancies, a number of scaffolds determined to be "unmappable" on the WashU map could be ordered relative to the anchored scaffolds

with GM99. These scaffolds were termed "ordered scaffolds." We found that 13.9% of the assembly could be ordered by these additional methods, and thus 84.0% of the genome was ordered unambiguously.

Next, all scaffolds that could be placed, but not ordered, between anchors were assigned to the interval between the anchored scaffolds and were deemed to be "bounded" between them. For example, small scaffolds having STS hits from the same GeneMap bin or hitting the same BAC cannot be ordered relative to each other, but can be assigned a placement boundary relative to other anchored or ordered scaffolds. The remaining scaffolds either had no localization information, conflicting information, or could only be assigned to a generic chromosome location. Using the above approaches, ~98% of the genome was anchored, ordered, or bounded.

Finally, we assigned a location for each scaffold placed on the chromosome by spreading out the scaffolds per chromosome. We assumed that the remaining unmapped scaffolds, constituting 2% of the genome, were distributed evenly across the genome. By dividing the sum of unmapped scaffold lengths with the sum of the number of mapped scaffolds, we arrived at an estimate of interscaffold gap of 1483 bp. This gap was used to separate all the scaffolds on each chromosome and to assign an offset in the chromosome.

During the scaffold-mapping effort, we encountered many problems that resulted in additional quality assessment and validation analysis. At least 978 (3% of 33,173) BACs were believed to have sequence data from more than one location in the genome (*47*). This is consistent with the bactig chimerism analysis reported above in the Assembly Strategies section. These BACs could not be assigned to unique positions within the CSA assembly and thus could not be used for ordering scaffolds. Likewise, it was not always possible to assign STSs to unique locations in the assembly because of genome duplications, repetitive elements, and pseudogenes.

Because of the time required for an exhaustive search for a perfect overlap, CSA generated 21,607 intrascaffold gaps where the mate-pair data suggested that the contigs should overlap, but no overlap was found. These gaps were defined as a fixed 50 bp in length and make up 18.6% of the total 116,442 gaps in the CSA assembly.

We chose not to use the order of exons implied in cDNA or EST data as a way of ordering scaffolds. The rationale for not using this data was that doing so would have biased certain regions of the assembly by rearranging scaffolds to fit the transcript data and made validation of both the assembly and gene definition processes more difficult.

2.7 Assembly and validation analysis

We analyzed the assembly of the genome from the perspectives of completeness (amount of coverage of the genome) and correctness (the structural accuracy of the order and orientation and the consensus sequence of the assembly).

Completeness. Completeness is defined as the percentage of the euchromatic sequence represented in the assembly. This cannot be known with absolute certainty until the euchromatin sequence has been completed. However, it is possible to estimate completeness on the basis of (i) the estimated sizes of intrascaffold gaps; (ii) coverage of the two published chromosomes, 21 and 22 (*48*, *49*); and (iii) analysis of the percentage of an independent set of random sequences (STS markers) contained in the assembly. The whole-genome libraries contain heterochromatic sequence and, although no attempt has been made to assemble it, there may be instances of unique sequence embedded in regions of heterochromatin as were observed in *Drosophila* (*50*, *51*).

The sequences of human chromosomes 21 and 22 have been completed to high quality and published (*48*, *49*). Although this sequence served as input to the assembler, the finished sequence was shredded into a shotgun data set so that the assembler had the opportunity to assemble it differently from the original sequence in the case of structural polymorphisms or assembly errors in the BAC data. In particular, the assembler must be able to resolve repetitive elements at the scale of components (generally multimegabase in size), and so this comparison reveals the level to which the assembler resolves repeats. In certain areas, the assembly structure differs from the published versions of chromosomes 21 and 22 (see below). The consequence of the flexibility to assemble "finished" sequence differently on the basis of Celera data resulted in an assembly with more segments than the chromosome 21 and 22 sequences. We examined the reasons why there are more gaps in the Celera sequence than in chromosomes 21 and 22 and expect that they may be typical of gaps in other regions of the genome. In the Celera assembly, there are 25 scaffolds, each containing at least 10 kb of sequence, that collectively span 94.3% of chromosome 21. Sixty-two scaffolds span 95.7% of chromosome 22. The total length of the gaps remaining in the Celera assembly for these two chromosomes is 3.4 Mbp. These gap sequences were analyzed by RepeatMasker and by searching against the entire genome assembly (*52*). About 50% of the gap sequence consisted of common repetitive elements identified by RepeatMasker; more than half of the remainder was lower copy number repeat elements.

A more global way of assessing completeness is to measure the content of an independent set of sequence data in the assembly. We compared 48,938 STS markers from Genemap99 (*51*) to the scaffolds. Because these markers were not used in the assembly processes, they provided a truly independent measure of completeness. ePCR (*53*) and BLAST (*54*) were used to locate STSs on the assembled genome. We found 44,524 (91%) of the STSs in the mapped genome. An additional 2648 markers (5.4%) were found by searching the unassembled data or "chaff." We identified 1283 STS markers (2.6%) not found in either Celera sequence or BAC data as of September 2000, raising the possibility that these markers may not be of human origin. If that were the case, the Celera assembled sequence would represent 93.4% of the human genome and the unassembled data 5.5%, for a total of 98.9% coverage. Similarly, we compared CSA against 36,678 TNG radiation hybrid markers (*55a*) using the same method. We found that 32,371 markers (88%) were located in the mapped CSA scaffolds, with 2055 markers (5.6%) found in the remainder. This gave a 94% coverage of the genome through another genome-wide survey.

Correctness. Correctness is defined as the structural and sequence accuracy of the assembly. Because the source sequences for the Celera data and the GenBank data are from different individuals, we could not directly compare the consensus sequence of the as-

Table 4. Summary of scaffold mapping. Scaffolds were mapped to the genome with different levels of confidence (anchored scaffolds have the highest confidence; unmapped scaffolds have the lowest). Anchored scaffolds were consistently ordered by the WashU BAC map and GM99. Ordered scaffolds were consistently ordered by at least one of the following: the WashU BAC map, GM99, or component tiling path. Bounded scaffolds had order conflicts between at least two of the external maps, but their placements were adjacent to a neighboring anchored or ordered scaffold. Unmapped scaffolds had, at most, a chromosome assignment. The scaffold subcategories are given below each category.

Mapped scaffold category	Number	Length (bp)	% Total length
Anchored	1,526	1,860,676,676	70
Oriented	1,246	1,852,088,645	70
Unoriented	280	8,588,031	0.3
Ordered	2,001	369,235,857	14
Oriented	839	329,633,166	12
Unoriented	1,162	39,602,691	2
Bounded	38,241	368,753,463	14
Oriented	7,453	274,536,424	10
Unoriented	30,788	94,217,039	4
Unmapped	11,823	55,313,737	2
Known chromosome	281	2,505,844	0.1
Unknown chromosome	11,542	52,807,893	2

sembly against other finished sequence for determining sequencing accuracy at the nucleotide level, although this has been done for identifying polymorphisms as described in Section 6. The accuracy of the consensus sequence is at least 99.96% on the basis of a statistical estimate derived from the quality values of the underlying reads.

The structural consistency of the assembly can be measured by mate-pair analysis. In a correct assembly, every mated pair of sequencing reads should be located on the consensus sequence with the correct separation and orientation between the pairs. A pair is termed "valid" when the reads are in the correct orientation and the distance between them is within the mean ± 3 standard deviations of the distribution of insert sizes of the library from which the pair was sampled. A pair is termed "misoriented" when the reads are not correctly oriented, and is termed "misseparated" when the distance between the reads is not in the correct range but the reads are correctly oriented. The mean ± the standard deviation of each library used by the assembler was determined as described above. To validate these, we examined all reads mapped to the finished sequence of chromosome 21 (*48*) and determined how many incorrect mate pairs there were as a result of laboratory tracking errors and chimerism (two different segments of the genome cloned into the same plasmid), and how tight the distribution of insert sizes was for those that were correct (Table 5). The standard deviations for all Celera libraries were quite small, less than 15% of the insert length, with the exception of a few 50-kbp libraries. The 2- and 10-kbp libraries contained less than 2% invalid mate pairs, whereas the 50-kbp libraries were somewhat higher (~10%). Thus, although the mate-pair information was not perfect, its accuracy was such that measuring valid, misoriented, and misseparated pairs with respect to a given assembly was deemed to be a reliable instrument for validation purposes, especially when several mate pairs confirm or deny an ordering.

The clone coverage of the genome was 39×, meaning that any given base pair was, on average, contained in 39 clones or, equivalently, spanned by 39 mate-paired reads. Areas of low clone coverage or areas with a high proportion of invalid mate pairs would indicate potential assembly problems. We computed the coverage of each base in the assembly by valid mate pairs (Table 6). In summary, for scaffolds >30 kbp in length, less than 1% of the Celera assembly was in regions of less than 3× clone coverage. Thus, more than 99% of the assembly, including order and orientation, is strongly supported by this measure alone.

We examined the locations and number of all misoriented and misseparated mates. In addition to doing this analysis on the CSA assembly (as of 1 October 2000), we also performed a study of the PFP assembly as of 5 September 2000 (*30*, *55b*). In this latter case, Celera mate pairs had to be mapped to the PFP assembly. To avoid mapping errors due to high-fidelity repeats, the only pairs mapped were those for which both reads matched at only one location with less than 6% differences. A threshold was set such that sets of five or more simultaneously invalid mate pairs indicated a potential breakpoint, where the construction of the two assemblies differed. The graphic comparison of the CSA chromosome 21 assembly with the published sequence (Fig. 6A) serves as a validation of this methodology. Blue tick marks in the panels indicate breakpoints. There were a similar (small) number of breakpoints on both chromosome sequences. The exception was 12 sets of scaffolds in the Celera assembly (a total of 3% of the chromosome length in 212 single-contig scaffolds) that were mapped to the wrong positions because they were too small to be mapped reliably. Figures 6 and 7 and Table 6 illustrate the mate-pair differences and breakpoints between the two assemblies. There was a higher percentage of misoriented and misseparated mate pairs in the large-insert libraries (50 kbp and BAC ends) than in the small-insert libraries in both assemblies (Table 6). The large-insert libraries are more likely to identify discrepancies simply because they span a larger segment of the genome. The graphic comparison between the two assemblies for chromosome 8 (Fig. 6, B and C) shows that there are many

Table 5. Mate-pair validation. Celera fragment sequences were mapped to the published sequence of chromosome 21. Each mate pair uniquely mapped was evaluated for correct orientation and placement (number of mate pairs tested). If the two mates had incorrect relative orientation or placement, they were considered invalid (number of invalid mate pairs).

		Chromosome 21						Genome		
Library type	Library no.	Mean insert size (bp)	SD (bp)	SD/ mean (%)	No. of mate pairs tested	No. of invalid mate pairs	% invalid	Mean insert size (bp)	SD (bp)	SD/ mean (%)
2 kbp	1	2,081	106	5.1	3,642	38	1.0	2,082	90	4.3
	2	1,913	152	7.9	28,029	413	1.5	1,923	118	6.1
	3	2,166	175	8.1	4,405	57	1.3	2,162	158	7.3
10 kbp	4	11,385	851	7.5	4,319	80	1.9	11,370	696	6.1
	5	14,523	1,875	12.9	7,355	156	2.1	14,142	1,402	9.9
	6	9,635	1,035	10.7	5,573	109	2.0	9,606	934	9.7
	7	10,223	928	9.1	34,079	399	1.2	10,190	777	7.6
50 kbp	8	64,888	2,747	4.2	16	1	6.3	65,500	5,504	8.4
	9	53,410	5,834	10.9	914	170	18.6	53,311	5,546	10.4
	10	52,034	7,312	14.1	5,871	569	9.7	51,498	6,588	12.8
	11	52,282	7,454	14.3	2,629	213	8.1	52,282	7,454	14.3
	12	46,616	7,378	15.8	2,153	215	10.0	45,418	9,068	20.0
	13	55,788	10,099	18.1	2,244	249	11.1	53,062	10,893	20.5
	14	39,894	5,019	12.6	199	7	3.5	36,838	9,988	27.1
BES	15	48,931	9,813	20.1	144	10	6.9	47,845	4,774	10.0
	16	48,130	4,232	8.8	195	14	7.2	47,924	4,581	9.6
	17	106,027	27,778	26.2	330	16	4.8	152,000	26,600	17.5
	18	160,575	54,973	34.2	155	8	5.2	161,750	27,000	16.7
	19	164,155	19,453	11.9	642	44	6.9	176,500	19,500	11.05
Sum					102,894	2,768 (mean = 2.7)	2.7			

more breakpoints for the PFP assembly than for the Celera assembly. Figure 7 shows the breakpoint map (blue tick marks) for both assemblies of each chromosome in a side-by-side fashion. The order and orientation of Celera's assembly shows substantially fewer breakpoints except on the two finished chromosomes. Figure 7 also depicts large gaps (>10 kbp) in both assemblies as red tick marks. In the CSA assembly, the size of all gaps have been estimated on the basis of the mate-pair data. Breakpoints can be caused by structural polymorphisms, because the two assemblies were derived from different human genomes. They also reflect the unfinished nature of both genome assemblies.

3 Gene Prediction and Annotation

Summary. To enumerate the gene inventory, we developed an integrated, evidence-based approach named Otto. The evidence used to increase the likelihood of identifying genes includes regions conserved between the mouse and human genomes, similarity to ESTs or other mRNA-derived data, or similarity to other proteins. A comparison of Otto (combined Otto-RefSeq and Otto homology) with Genscan, a standard gene-prediction algorithm, showed greater sensitivity (0.78 versus 0.50) and specificity (0.93 versus 0.63) of Otto in the ability to define gene structure. Otto-predicted genes were complemented with a set of genes from three gene-prediction programs that exhibited weaker, but still significant, evidence that they may be expressed. Conservative criteria, requiring at least two lines of evidence, were used to define a set of 26,383 genes with good confidence that were used for more detailed analysis presented in the subsequent sections. Extensive manual curation to establish precise characterization of gene structure will be necessary to improve the results from this initial computational approach.

3.1 Automated gene annotation

A gene is a locus of cotranscribed exons. A single gene may give rise to multiple transcripts, and thus multiple distinct proteins with multiple functions, by means of alternative splicing and alternative transcription initiation and termination sites. Our cells are able to discern within the billions of base pairs of the genomic DNA the signals for initiating transcription and for splicing together exons separated by a few or hundreds of thousands of base pairs. The first step in characterizing the genome is to define the structure of each gene and each transcription unit.

The number of protein-coding genes in mammals has been controversial from the outset. Initial estimates based on reassociation data placed it between 30,000 to 40,000, whereas later estimates from the brain were >100,000 (*56*). More recent data from both the corporate and public sectors, based on extrapolations from EST, CpG island, and transcript density–based extrapolations, have not reduced this variance. The highest recent number of 142,634 genes emanates from a report from Incyte Pharmaceuticals, and is based on a combination of EST data and the association of ESTs with CpG islands (*57*). In stark contrast are three quite different, and much lower estimates: one of ~35,000 genes derived with genome-wide EST data and sampling procedures in conjunction with chromosome 22 data (*58*); another of 28,000 to 34,000 genes derived with a comparative methodology involving sequence conservation between humans and the puffer fish *Tetraodon nigroviridis* (*59*); and a figure of 35,000 genes, which was derived simply by extrapolating from the density of 770 known and predicted genes in the 67 Mbp of chromosomes 21 and 22, to the approximately 3-Gbp euchromatic genome.

The problem of computational identification of transcriptional units in genomic DNA sequence can be divided into two phases. The first is to partition the sequence into segments that are likely to correspond to individual genes. This is not trivial and is a weakness of most de novo gene-finding algorithms. It is also critical to determining the number of genes in the human gene inventory. The second challenge is to construct a gene model that reflects the probable structure of the transcript(s) encoded in the region. This can be done with reasonable accuracy when a full-length cDNA has been sequenced or a highly homologous protein sequence is known. De novo gene prediction, although less accurate, is the only way to find genes that are not represented by homologous proteins or ESTs. The following section describes the methods we have developed to address these problems for the prediction of protein-coding genes.

We have developed a rule-based expert system, called Otto, to identify and characterize genes in the human genome (*60*). Otto attempts to simulate in software the process that a human annotator uses to identify a gene and refine its structure. In the process of annotating a region of the genome, a human curator examines the evidence provided by the computational pipeline (described below) and examines how various types of evidence relate to one another. A curator puts different levels of confidence in different types of evidence and looks for certain patterns of evidence to support gene annotation. For example, a curator may examine homology to a number of ESTs and evaluate whether or not they can be connected into a longer, virtual mRNA. The curator would also evaluate the strength of the similarity and the contiguity of the match, in essence asking whether any ESTs cross splice-junctions and whether the edges of putative exons have consensus splice sites. This kind of manual annotation process was used to annotate the *Drosophila* genome.

The Otto system can promote observed evidence to a gene annotation in one of two ways. First, if the evidence includes a high-quality match to the sequence of a known gene [here defined as a human gene represented in a curated subset of the RefSeq database (*61*)], then Otto can promote this to a gene annotation. In the second method, Otto evaluates a broad spectrum of evidence and determines if this evidence is adequate to support promotion to a gene annotation. These processes are described below.

Initially, gene boundaries are predicted on the basis of examination of sets of overlapping protein and EST matches generated by a computational pipeline (*62*). This pipeline searches the scaffold sequences against protein, EST, and genome-sequence databases to define regions of sequence similarity and runs three de novo gene-prediction programs.

To identify likely gene boundaries, regions of the genome were partitioned by Otto on the basis of sequence matches identified by BLAST. Each of the database sequences matched in the region under analysis was compared by an algorithm that takes into account both coordinates of the matching sequence, as well as the sequence type (e.g. protein, EST, and so forth). The results were used to group the matches into bins of related sequences that may define a gene and identi-

Table 6. Genome-wide mate pair analysis of compartmentalized shotgun (CSA) and PFP assemblies.*

Genome library	CSA			PFP		
	% valid	% mis-oriented	% mis-separated†	% valid	% mis-oriented	% mis-separated†
2 kbp	98.5	0.6	1.0	95.7	2.0	2.3
10 kbp	96.7	1.0	2.3	81.9	9.6	8.6
50 kbp	93.9	4.5	1.5	64.2	22.3	13.5
BES	94.1	2.1	3.8	62.0	19.3	18.8
Mean	97.4	1.0	1.6	87.3	6.8	5.9

*Data for individual chromosomes can be found in Web fig. 3 on *Science* Online at www.sciencemag.org/cgi/content/full/291/5507/1304/DC1. †Mates are misseparated if their distance is >3 SD from the mean library size.

gene boundaries. During this process, multiple hits to the same region were collapsed to a coherent set of data by tracking the coverage of a region. For example, if a group of bases was represented by multiple overlapping ESTs, the union of these regions matched by the set of ESTs on the scaffold was marked as being supported by EST evidence. This resulted in a series of "gene bins," each of which was believed to contain a single gene. One weakness of this initial implementation of the algorithm was in predicting gene boundaries in regions of tandemly duplicated genes. Gene clusters frequently resulted in homologous neighboring genes being joined together, resulting in an annotation that artificially concatenated these gene models.

Next, known genes (those with exact matches of a full-length cDNA sequence to the genome) were identified, and the region corresponding to the cDNA was annotated as a predicted transcript. A subset of the curated human gene set RefSeq from the National Center for Biotechnology Information (NCBI) was included as a data set searched in the computational pipeline. If a RefSeq transcript matched the genome assembly for at least 50% of its length at >92% identity, then the SIM4 (*63*) alignment of the RefSeq transcript to the region of the genome under analysis was promoted to the status of an Otto annotation. Because the genome sequence has gaps and sequence errors such as frameshifts, it was not always possible to predict a transcript that agrees precisely with the experimentally determined cDNA sequence. A total of 6538 genes in our inventory were identified and transcripts predicted in this way.

Regions that have a substantial amount of sequence similarity, but do not match known genes, were analyzed by that part of the Otto system that uses the sequence similarity information to predict a transcript. Here, Otto

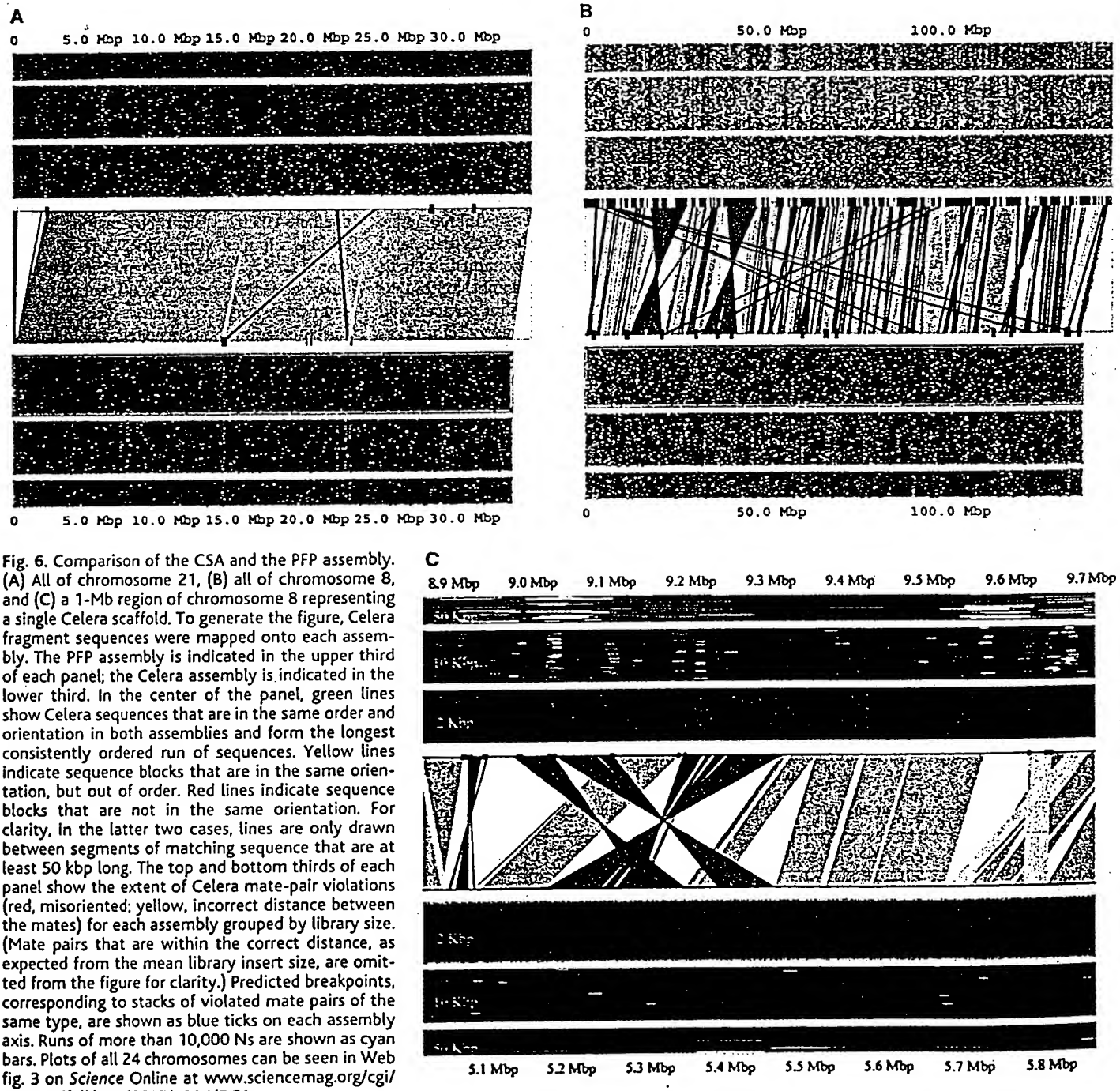


Fig. 6. Comparison of the CSA and the PFP assembly. (A) All of chromosome 21, (B) all of chromosome 8, and (C) a 1-Mb region of chromosome 8 representing a single Celera scaffold. To generate the figure, Celera fragment sequences were mapped onto each assembly. The PFP assembly is indicated in the upper third of each panel; the Celera assembly is indicated in the lower third. In the center of the panel, green lines show Celera sequences that are in the same order and orientation in both assemblies and form the longest consistently ordered run of sequences. Yellow lines indicate sequence blocks that are in the same orientation, but out of order. Red lines indicate sequence blocks that are not in the same orientation. For clarity, in the latter two cases, lines are only drawn between segments of matching sequence that are at least 50 kbp long. The top and bottom thirds of each panel show the extent of Celera mate-pair violations (red, misoriented; yellow, incorrect distance between the mates) for each assembly grouped by library size. (Mate pairs that are within the correct distance, as expected from the mean library insert size, are omitted from the figure for clarity.) Predicted breakpoints, corresponding to stacks of violated mate pairs of the same type, are shown as blue ticks on each assembly axis. Runs of more than 10,000 Ns are shown as cyan bars. Plots of all 24 chromosomes can be seen in Web fig. 3 on *Science* Online at www.sciencemag.org/cgi/content/full/291/5507/1304/DC1.

evaluates evidence generated by the computational pipeline, corresponding to conservation between mouse and human genomic DNA, similarity to human transcripts (ESTs and cDNAs), similarity to rodent transcripts (ESTs and cDNAs), and similarity of the translation of human genomic DNA to known proteins to predict potential genes in the human genome. The sequence from the region of genomic DNA contained in a gene bin was extracted, and the subsequences supported by any homology evidence were marked (plus 100

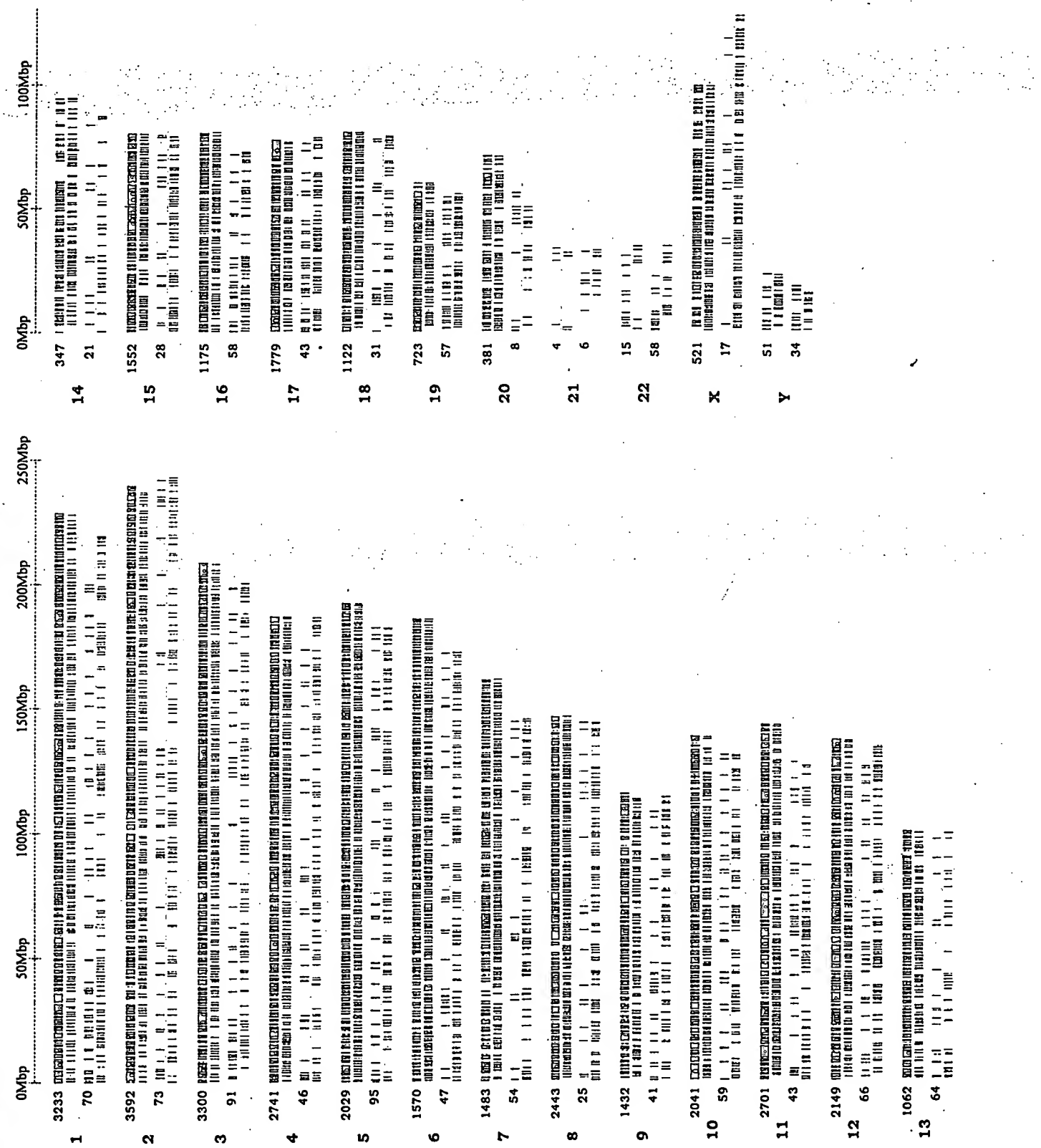


Fig. 7. Schematic view of the distribution of breakpoints and large gaps on all chromosomes. For each chromosome, the upper pair of lines represent the PFP assembly, and the lower pair of lines represent Celera's assembly. Blue tick marks represent breakpoints, whereas red tick marks represent a gap of larger than 10,000 bp. The number of breakpoints per chromosome is indicated in black, and the chromosome numbers in red.

bases flanking these regions). The other bases in the region, those not covered by any homology evidence, were replaced by N's. This sequence segment, with high confidence regions represented by the consensus genomic sequence and the remainder represented by N's, was then evaluated by Genscan to see if a consistent gene model could be generated. This procedure simplified the gene-prediction task by first establishing the boundary for the gene (not a strength of most gene-finding algorithms), and by eliminating regions with no supporting evidence. If Genscan returned a plausible gene model, it was further evaluated before being promoted to an "Otto" annotation. The final Genscan predictions were often quite different from the prediction that Genscan returned on the same region of native genomic sequence. A weakness of using Genscan to refine the gene model is the loss of valid, small exons from the final annotation.

The next step in defining gene structures based on sequence similarity was to compare each predicted transcript with the homology-based evidence that was used in previous steps to evaluate the depth of evidence for each exon in the prediction. Internal exons were considered to be supported if they were covered by homology evidence to within ±10 bases of their edges. For first and last exons, the internal edge was required to be within 10 bases, but the external edge was allowed greater latitude to allow for 5′ and 3′ untranslated regions (UTRs). To be retained, a prediction for a multi-exon gene must have evidence such that the total number of "hits," as defined above, divided by the number of exons in the prediction must be >0.66 or must correspond to a RefSeq sequence. A single-exon gene must be covered by at least three supporting hits (±10 bases on each side), and these must cover the complete predicted open reading frame. For a single-exon gene, we also required that the Genscan prediction include both a start and a stop codon. Gene models that did not meet these criteria were disregarded, and those that passed were promoted to Otto predictions. Homology-based Otto predictions do not contain 3′ and 5′ untranslated sequence. Although three de novo gene-finding programs [GRAIL, Genscan, and FgenesH (*63*)] were run as part of the computational analysis, the results of these programs were not directly used in making the Otto predictions. Otto predicted 11,226 additional genes by means of sequence similarity.

Table 7. Sensitivity and specificity of Otto and Genscan. Sensitivity and specificity were calculated by first aligning the prediction to the published RefSeq transcript, tallying the number (*N*) of uniquely aligned RefSeq bases. Sensitivity is the ratio of *N* to the length of the published RefSeq transcript. Specificity is the ratio of *N* to the length of the prediction. All differences are significant (Tukey HSD; $P < 0.001$).

Method	Sensitivity	Specificity
Otto (RefSeq only)*	0.939	0.973
Otto (homology)†	0.604	0.884
Genscan	0.501	0.633

*Refers to those annotations produced by Otto using only the Sim4-polished RefSeq alignment rather than an evidence-based Genscan prediction. †Refers to those annotations produced by supplying all available evidence to Genscan.

3.2 Otto validation

To validate the Otto homology-based process and the method that Otto uses to define the structures of known genes, we compared transcripts predicted by Otto with their corresponding (and presumably correct) transcript from a set of 4512 RefSeq transcripts for which there was a unique SIM4 alignment (Table 7). In order to evaluate the relative performance of Otto and Genscan, we made three comparisons. The first involved a determination of the accuracy of gene models predicted by Otto with only homology data other than the corresponding RefSeq sequence (Otto homology in Table 7). We measured the sensitivity (correctly predicted bases divided by the total length of the cDNA) and specificity (correctly predicted bases divided by the sum of the correctly and incorrectly predicted bases). Second, we examined the sensitivity and specificity of the Otto predictions that were made solely with the RefSeq sequence, which is the process that Otto uses to annotate known genes (Otto-RefSeq). And third, we determined the accuracy of the Genscan predictions corresponding to these RefSeq sequences. As expected, the alignment method (Otto-RefSeq) was the most accurate, and Otto-homology performed better than Genscan by both criteria. Thus, 6.1% of true RefSeq nucleotides were not represented in the Otto-refseq annotations and 2.7% of the nucleotides in the Otto-RefSeq transcripts were not contained in the original RefSeq transcripts. The discrepancies could come from legitimate differences between the Celera assembly and the RefSeq transcript due to polymorphisms, incomplete or incorrect data in the Celera assembly, errors introduced by Sim4 during the alignment process, or the presence of alternatively spliced forms in the data set used for the comparisons.

Because Otto uses an evidence-based approach to reconstruct genes, the absence of experimental evidence for intervening exons may inadvertantly result in a set of exons that cannot be spliced together to give rise to a transcript. In such cases, Otto may "split genes" when in fact all the evidence should be combined into a single transcript. We also examined the tendency of these methods to incorrectly split gene predictions. These trends are shown in Fig. 8. Both RefSeq and homology-based predictions by Otto split known genes into fewer segments than Genscan alone.

3.3 Gene number

Recognizing that the Otto system is quite conservative, we used a different gene-prediction strategy in regions where the homology evidence was less strong. Here the results of de novo gene predictions were used. For these genes, we insisted that a predicted transcript have at least two of the following types of evidence to be included in the gene set for further analysis: protein, human EST, rodent EST, or mouse genome fragment matches. This final class of predicted genes is a subset of the predictions made by the three gene-finding programs that were used in the computational pipeline. For these, there was not sufficient sequence similarity information for Otto to attempt to predict a gene structure. The three de novo gene-finding programs resulted in about 155,695 predictions, of which ~76,410 were nonredundant (nonoverlapping with one another). Of these, 57,935 did not overlap known genes or predictions made by Otto. Only 21,350 of the gene predictions that did not overlap Otto predictions were partially supported by at least one type of sequence similarity evidence, and 8619 were partially supported by two types of evidence (Table 8).

The sum of this number (21,350) and the number of Otto annotations (17,764), 39,114, is near the upper limit for the human gene complement. As seen in Table 8, if the requirement for other supporting evidence is made more stringent, this number drops rapidly so that demanding two types of evidence reduces the total gene number to 26,383 and demanding three types reduces it to ~23,000. Requiring that a prediction be supported by all four categories of evidence is too stringent because it would eliminate genes that encode novel proteins (members of currently undescribed protein families). No correction for pseudogenes has been made at this point in the analysis.

In a further attempt to identify genes that were not found by the autoannotation process or any of the de novo gene finders, we examined regions outside of gene predictions that were similar to the EST sequence, and where the EST matched the genomic sequence across a splice junction. After correcting for potential 3′ UTRs of predicted genes, about 2500 such regions remained. Addition of a requirement for at least one of the following evidence types—homology to mouse genomic sequence fragments, rodent ESTs, or cDNAs—or similarity to a known protein reduced this number to 1010. Adding this to the numbers from the previous paragraph would give us estimates of about 40,000, 27,000, and 24,000 potential genes in the human genome, depending on the stringency of evidence considered. Table 8 illustrates the number of genes and presents the degree of

confidence based on the supporting evidence. Transcripts encoded by a set of 26,383 genes were assembled for further analysis. This set includes the 6538 genes predicted by Otto on the basis of matches to known genes, 11,226 transcripts predicted by Otto based on homology evidence, and 8619 from the subset of transcripts from de novo gene-prediction programs that have two types of supporting evidence. The 26,383 genes are illustrated along chromosome diagrams in Fig. 1. These are a very preliminary set of annotations and are subject to all the limitations of an automated process. Considerable refinement is still necessary to improve the accuracy of these transcript predictions. All the predictions and descriptions of genes and the associated evidence that we present are the product of completely computational processes, not expert curation. We have attempted to enumerate the genes in the human genome in such a way that we have different levels of confidence based on the amount of supporting evidence: known genes, genes with good protein or EST homology evidence, and de novo gene predictions confirmed by modest homology evidence.

3.4 Features of human gene transcripts

We estimate the average span for a "typical" gene in the human DNA sequence to be about 27,894 bases. This is based on the average span covered by RefSeq transcripts, used because it represents our highest confidence set.

The set of transcripts promoted to gene annotations varies in a number of ways. As can be seen from Table 8 and Fig. 9, transcripts predicted by Otto tend to be longer, having on average about 7.8 exons, whereas those promoted from gene-prediction programs average about 3.7 exons. The largest number of exons that we have identified in a transcript is 234 in the titin mRNA. Table 8 compares the amounts of evidence that support the Otto and other predicted transcripts. For example, one can see that a typical Otto transcript has 6.99 of its 7.81 exons supported by protein homology evidence. As would be expected, the Otto transcripts generally have more support than do transcripts predicted by the de novo methods.

4 Genome Structure

Summary. This section describes several of the noncoding attributes of the assembled genome sequence and their correlations with the predicted gene set. These include an analysis of G+C content and gene density in the context of cytogenetic maps of the genome, an enumerative analysis of CpG islands, and a brief description of the genome-wide repetitive elements.

4.1 Cytogenetic maps

Perhaps the most obvious, and certainly the most visible, element of the structure of the genome is the banding pattern produced by Giemsa stain. Chromosomal banding studies have revealed that about 17% to 20% of the human chromosome complement consists of C-bands, or constitutive heterochromatin (*64*). Much of this heterochromatin is highly polymorphic and consists of different families of alpha satellite DNAs with various higher order repeat structures (*65*). Many chromosomes have complex inter- and intrachromosomal duplications present in pericentromeric regions (*66*). About 5% of the sequence reads were identified as alpha satellite sequences; these were not included in the assembly.

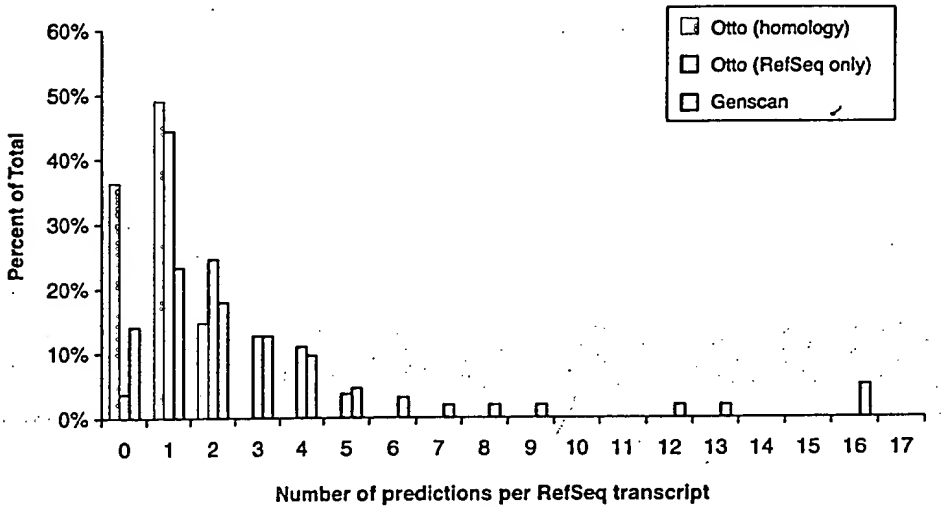


Fig. 8. Analysis of split genes resulting from different annotation methods. A set of 4512 Sim4-based alignments of RefSeq transcripts to the genomic assembly were chosen (see the text for criteria), and the numbers of overlapping Genscan, Otto (RefSeq only) annotations based solely on Sim4-polished RefSeq alignments, and Otto (homology) annotations (annotations produced by supplying all available evidence to Genscan) were tallied. These data show the degree to which multiple Genscan predictions and/or Otto annotations were associated with a single RefSeq transcript. The zero class for the Otto-homology predictions shown here indicates that the Otto-homology calls were made without recourse to the RefSeq transcript, and thus no Otto call was made because of insufficient evidence.

Table 8. Numbers of exons and transcripts supported by various types of evidence for Otto and de novo gene prediction methods. Highlighted cells indicate the gene sets analyzed in this paper (boldface, set of genes selected for protein analysis; italic, total set of accepted de novo predictions).

		Total	Types of evidence				No. of lines of evidence*			
			Mouse	Rodent	Protein	Human	≥1	≥2	≥3	≥4
Otto	Number of transcripts	17,969	17,065	14,881	15,477	16,374	17,968†	17,501	15,877	12,451
	Number of exons	141,218	111,174	89,569	108,431	118,869	140,710	127,955	99,574	59,804
De novo	Number of transcripts	58,032	14,463	5,094	8,043	9,220	*21,350*	8,619	4,947	1,904
	Number of exons	319,935	48,594	19,344	26,264	40,104	79,148	31,130	17,508	6,520
No. of exons per transcript	Otto	7.84	5.77	6.01	6.99	7.24	7.81	7.19	6.00	4.28
	De novo	5.53	3.17	3.80	3.27	4.36	3.7	3.56	3.42	3.16

*Four kinds of evidence (conservation in 3× mouse genomic DNA, similarity to human EST or cDNA, similarity to rodent EST or cDNA, and similarity to known proteins) were considered to support gene predictions from the different methods. The use of evidence is quite liberal, requiring only a partial match to a single exon of predicted transcript. †This number includes alternative splice forms of the 17,764 genes mentioned elsewhere in the text.

Examination of pericentromeric regions is ongoing.

The remaining ~80% of the genome, the euchromatic component, is divisible into G-, R-, and T-bands (*67*). These cytogenetic bands have been presumed to differ in their nucleotide composition and gene density, although we have been unable to determine precise band boundaries at the molecular level. T-bands are the most G+C- and gene-rich, and G-bands are G+C-poor (*68*). Bernardi has also offered a description of the euchromatin at the molecular level as long stretches of DNA of differing base composition, termed isochores (denoted L, H1, H2, and H3), which are >300 kbp in length (*69*). Bernardi defined the L (light) isochores as G+C-poor (<43%), whereas the H (heavy) isochores fall into three G+C-rich classes representing 24, 8, and 5% of the genome. Gene concentration has been claimed to be very low in the L isochores and 20-fold more enriched in the H2 and H3 isochores (*70*). By examining contiguous 50-kbp windows of G+C content across the assembly, we found that regions of G+C content >48% (H3 isochores) averaged 273.9 kbp in length, those with G+C content between 43 and 48% (H1+H2 isochores) averaged 202.8 kbp in length, and the average span of regions with <43% (L isochores) was 1078.6 kbp. The correlation between G+C content and gene density was also examined in 50-kbp windows along the assembled sequence (Table 9 and Figs. 10 and 11). We found that the density of genes was greater in regions of high G+C than in regions of low G+C content, as expected. However, the correlation between G+C content and gene density was not as skewed as previously predicted (*69*). A higher proportion of genes were located in the G+C-poor regions than had been expected.

Chromosomes 17, 19, and 22, which have a disproportionate number of H3-containing bands, had the highest gene density (Table 10). Conversely, of the chromosomes that we found to have the lowest gene density, X, 4, 18, 13, and Y, also have the fewest H3 bands. Chromosome 15, which also has few H3 bands, did not have a particularly low gene density in our analysis. In addition, chromosome 8, which we found to have a low gene density, does not appear to be unusual in its H3 banding.

How valid is Ohno's postulate (*71*) that mammalian genomes consist of oases of genes in otherwise essentially empty deserts? It appears that the human genome does indeed contain deserts, or large, gene-poor regions. If we define a desert as a region >500 kbp without a gene, then we see that 605 Mbp, or about 20% of the genome, is in deserts. These are not uniformly distributed over the various chromosomes. Gene-rich chromosomes 17, 19, and 22 have only about 12% of their collective 171 Mbp in deserts, whereas gene-poor chromosomes 4, 13, 18, and X have 27.5% of their 492 Mbp in deserts (Table 11). The apparent lack of predicted genes in these regions does not necessarily imply that they are devoid of biological function.

4.2 Linkage map

Linkage maps provide the basis for genetic analysis and are widely used in the study of the inheritance of traits and in the positional cloning of genes. The distance metric, centimorgans (cM), is based on the recombination rate between homologous chromosomes during meiosis. In general, the rate of recombination in females is greater than that in males, and this degree of map expansion is not uniform across the genome (*72*). One of the opportunities enabled by a nearly complete genome sequence is to produce the ultimate physical map, and to fully analyze its correspondence with two other maps that have been widely used in genome and genetic analysis: the linkage map and the cytogenetic map. This would close the loop between the mapping and sequencing phases of the genome project.

We mapped the location of the markers that constitute the Genethon linkage map to the genome. The rate of recombination, expressed as cM per Mbp, was calculated for 3-Mbp windows as shown in Table 12. Higher rates of recombination in the telomeric region of the chromosomes have been previously documented (*73*). From this mapping result, there is a difference of 4.99 between lowest rates and highest rates and the largest difference of 4.4 between males and females (4.99 to 0.47 on chromosome 16). This indicates that the variability in recombination rates among regions of the genome exceeds the differences in recombination rates between males and females. The human genome has recombination hotspots, where recombination rates vary fivefold or more over a space of 1 kbp, so the picture one gets of the magnitude of variability in recombination rate will depend on the size of the window

Table 9. Characteristics of G+C in isochores.

Isochore	G+C (%)	Fraction of genome		Fraction of genes	
		Predicted*	Observed	Predicted*	Observed
H3	>48	5	9.5	37	24.8
H1/H2	43–48	25	21.2	32	26.6
L	<43	67	69.2	31	48.5

*The predictions were based on Bernardi's definitions (*70*) of the isochore structure of the human genome.

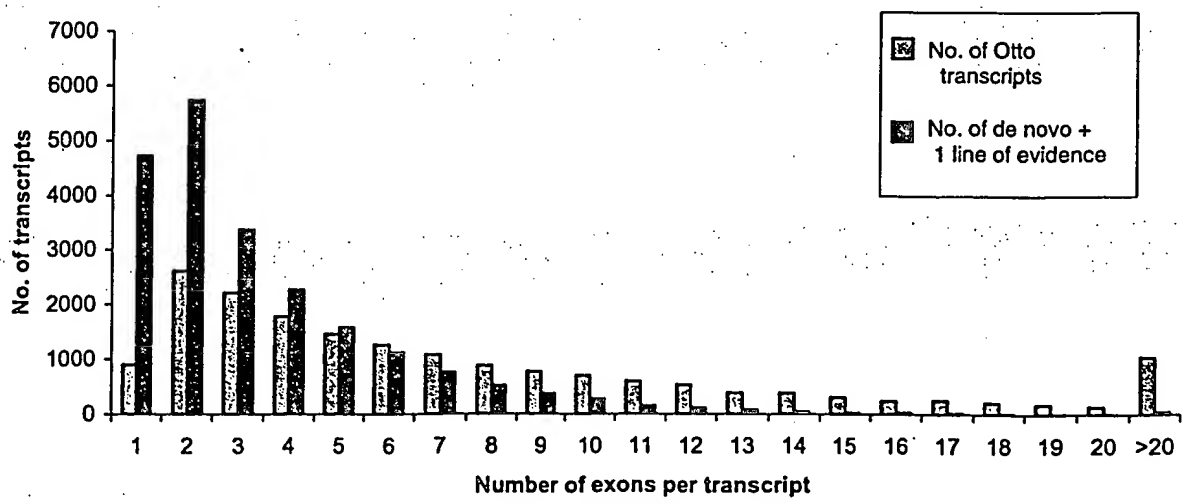


Fig. 9. Comparison of the number of exons per transcript between the 17,968 Otto transcripts and 21,350 de novo transcript predictions with at least one line of evidence that do not overlap with an Otto prediction. Both sets have the highest number of transcripts in the two-exon category, but the de novo gene predictions are skewed much more toward smaller transcripts. In the Otto set, 19.7% of the transcripts have one or two exons, and 5.7% have more than 20. In the de novo set, 49.3% of the transcripts have one or two exons, and 0.2% have more than 20.

examined. Unfortunately, too few meiotic crossovers have occurred in Centre d'Étude du Polymorphism Humain (CEPH) and other reference families to provide a resolution any finer than about 3 Mbp. The next challenge will be to determine a sequence basis of recombination at the chromosomal level. An accurate predictor for the rate for variation in recombination rates between any pair of markers would be extremely useful in designing markers to narrow a region of linkage, such as in positional cloning projects.

4.3 Correlation between CpG islands and genes

CpG islands are stretches of unmethylated DNA with a higher frequency of CpG dinucleotides when compared with the entire genome (*74*). CpG islands are believed to preferentially occur at the transcriptional start of genes, and it has been observed that most housekeeping genes have CpG islands at the 5′ end of the transcript (*75*, *76*). In addition, experimental evidence indicates that CpG island methylation is correlated with gene inactivation (*77*) and has been shown to be important during gene imprinting (*78*) and tissue-specific gene expression (*79*)

Experimental methods have been used that resulted in an estimate of 30,000 to 45,000 CpG islands in the human genome (*74*, *80*) and an estimate of 499 CpG islands on human chromosome 22 (*81*). Larsen *et al.* (*76*) and Gardiner-Garden and Frommer (*75*) used a computational method to identify CpG islands and defined them as regions of DNA of >200 bp that have a G+C content of >50% and a ratio of observed versus expected frequency of CG dinucleotide ≥0.6.

It is difficult to make a direct comparison of experimental definitions of CpG islands with computational definitions because computational methods do not consider the methylation state of cytosine and experimental methods do not directly select regions of high G+C content. However, we can determine the correlation of CpG island with gene starts, given a set of annotated genomic transcripts and the whole genome sequence. We have analyzed the publicly available annotation of chromosome 22, as well as using the entire human genome in our assembly and the computationally annotated genes. A variation of the CpG island computation was compared with Larsen *et al.* (*76*). The main differences are that we use a sliding window of 200 bp, consecutive windows are merged only if they overlap, and we recompute the CpG value upon merging, thus rejecting any potential island if it scores less than the threshold.

To compute various CpG statistics, we used two different thresholds of CG dinucleotide likelihood ratio. Besides using the original threshold of 0.6 (method 1), we used a higher threshold of CG dinucleotide likelihood ratio of 0.8 (method 2), which results in the number of CpG islands on chromosome 22 close to the number of annotated genes on this chromosome. The main results are summarized in Table 13. CpG islands computed with method 1 predicted only 2.6% of the CSA sequence as CpG, but 40% of the gene starts (start codons) are contained inside a CpG island. This is comparable to ratios reported by others (*82*). The last two rows of the table show the observed and expected average distance, respectively, of the closest CpG island from the first exon. The observed average closest CpG islands are smaller than the corresponding expected distances, confirming an association between CpG island and the first exon.

We also looked at the distribution of CpG island nucleotides among various sequence classes such as intergenic regions, introns, exons, and first exons. We computed the likelihood score for each sequence class as the ratio of the observed fraction of CpG island nucleotides in that sequence class and the expected fraction of CpG island nucleotides in that sequence class. The result of applying method 1 on CSA were scores of 0.89 for intergenic region, 1.2 for intron, 5.86 for exon, and 13.2 for first exon. The same trend was also found for chromosome 22 and after the application of a higher threshold (method 2) on both data sets. In sum, genome-wide analysis has extended earlier analysis and suggests a strong correlation between CpG islands and first coding exons.

4.4 Genome-wide repetitive elements

The proportion of the genome covered by various classes of repetitive DNA is presented in Table 14. We observed about 35% of the genome in these repeat classes, very similar to values reported previously (*83*). Repetitive sequence may be underrepresented in the Celera assembly as a result of incomplete repeat resolution, as discussed above. About 8% of the scaffold length is in gaps, and we expect that much of this is repetitive sequence. Chromosome 19 has the highest repeat density (57%), as well as the highest gene density (Table 10). Of interest, among the different classes of repeat elements, we observe a clear association of Alu elements and gene density, which was not observed between LINEs and gene density.

5 Genome Evolution

Summary. The dynamic nature of genome evolution can be captured at several levels. These include gene duplications mediated by RNA intermediates (retrotransposition) and segmental genomic duplications. In this section, we document the genome-wide occurrence of retrotransposition events generating functional (intronless paralogs) or inactive genes (pseudogenes). Genes involved in translational processes and nuclear regulation account for nearly 50% of all intronless paralogs and processed pseudogenes detected in our survey. We have also cataloged the extent of segmental genomic duplication and provide evidence for 1077 duplicated blocks covering 3522 distinct genes.

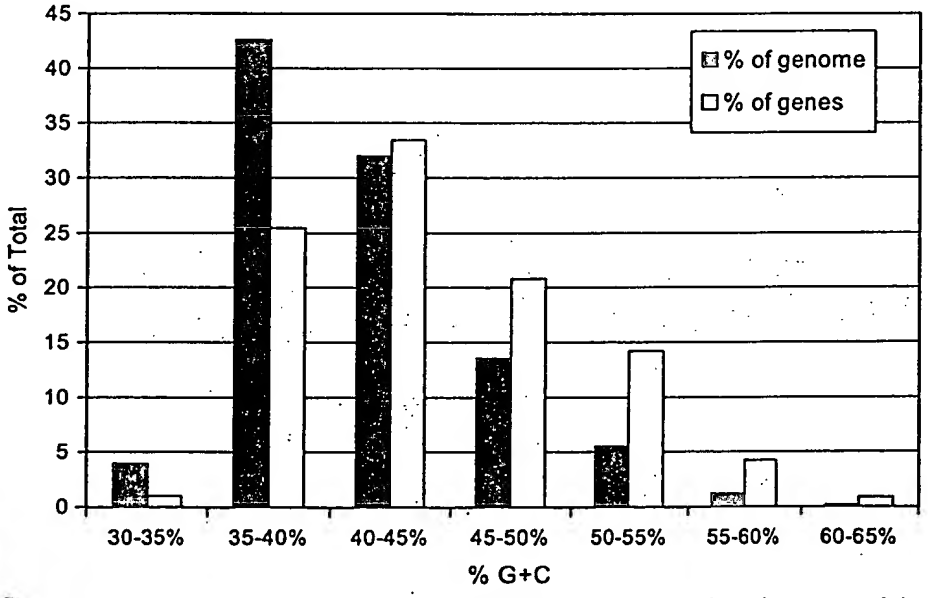


Fig. 10. Relation between G+C content and gene density. The blue bars show the percent of the genome (in 50-kbp windows) with the indicated G+C content. The percent of the total number of genes associated with each G+C bin is represented by the yellow bars. The graph shows that about 5% of the genome has a G+C content of between 50 and 55%, but that this portion contains nearly 15% of the genes.

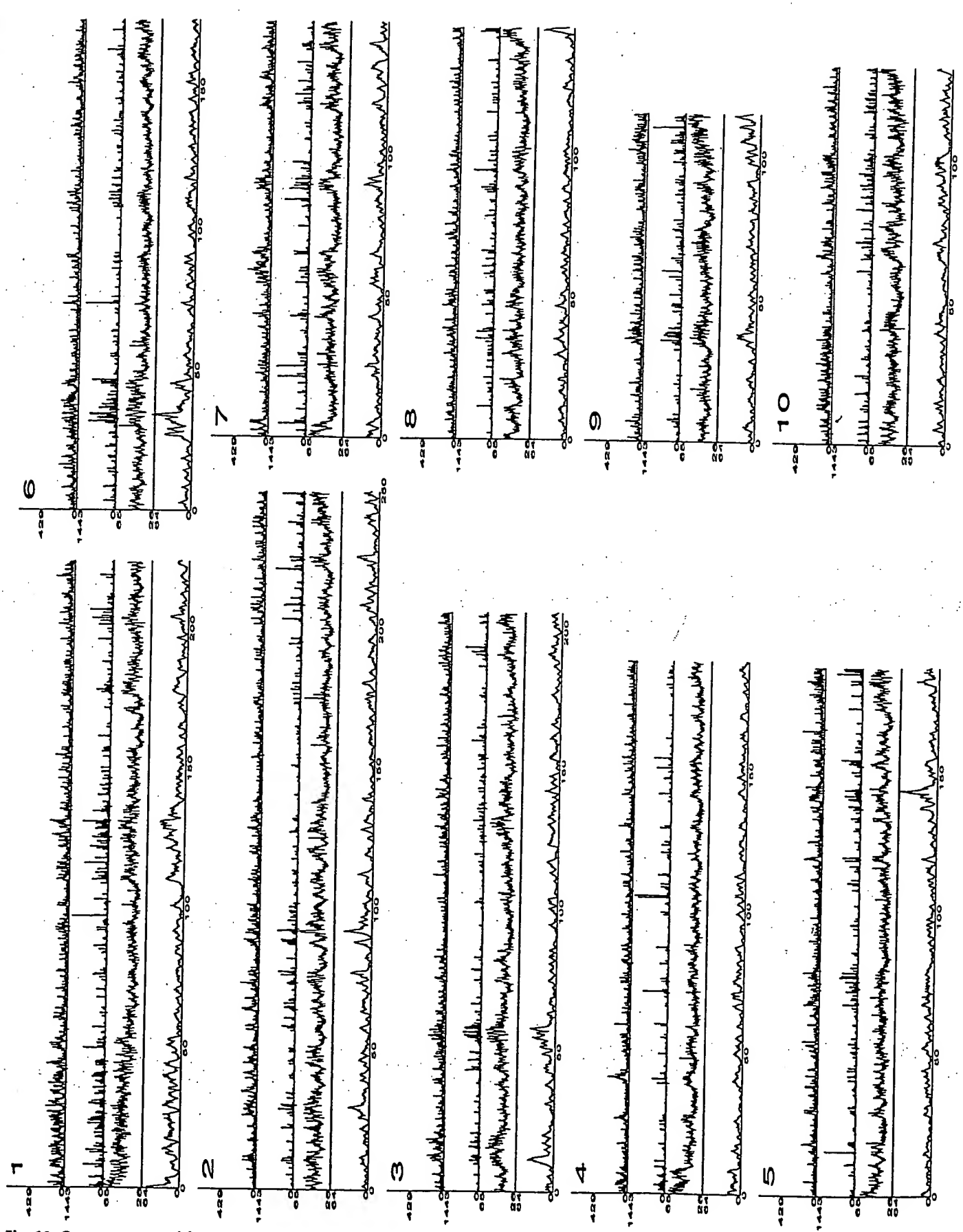


Fig. 11. Genome structural features.

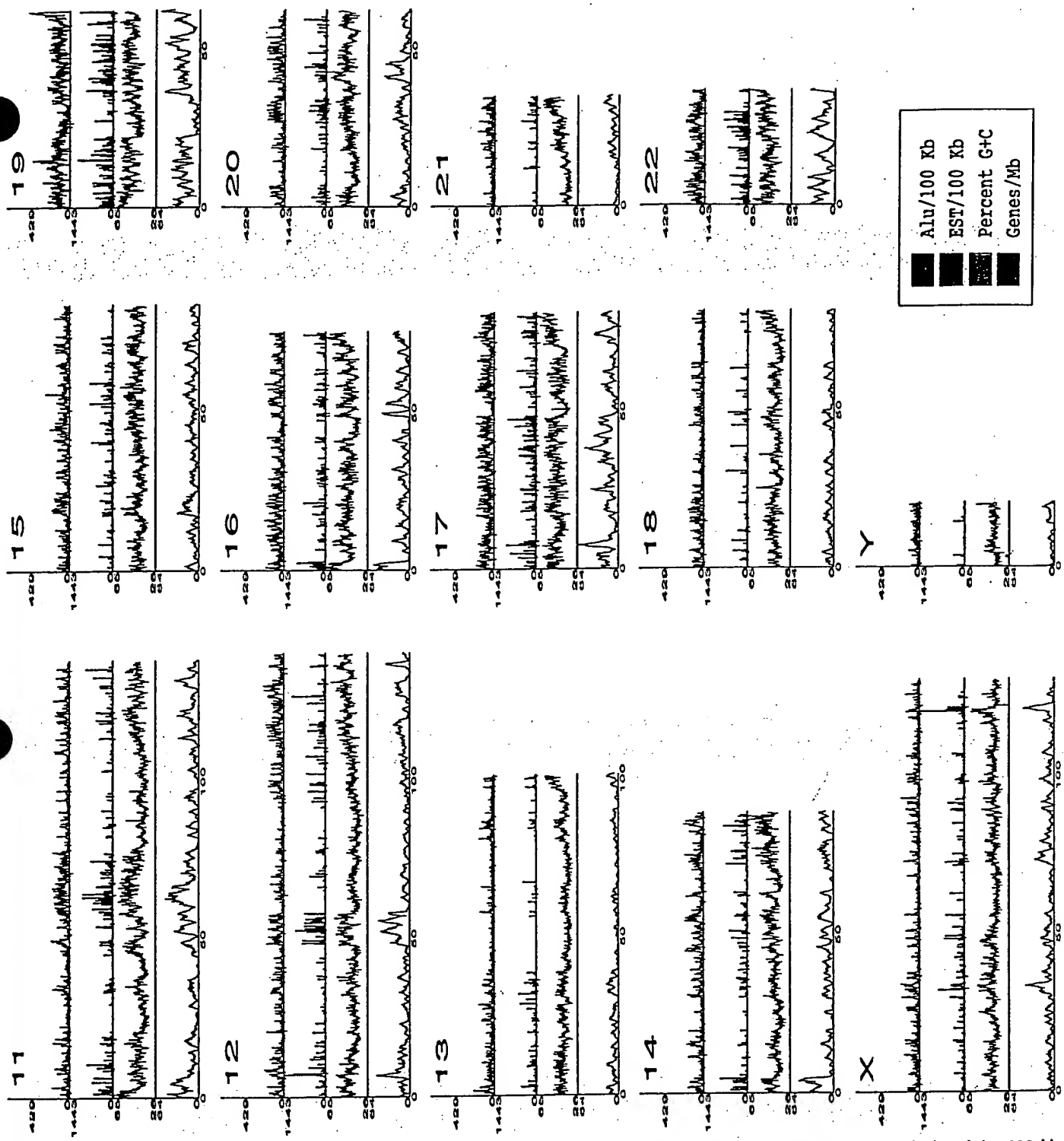


Fig. 11 (continued). Relation among gene density (orange), G+C content (green), EST density (blue), and Alu density (pink) along the lengths of each of the chromosomes. Gene density was calculated in 1-Mbp windows. The percent of G+C nucleotides was calculated in 100-kbp windows. The number of ESTs and Alu elements is shown per 100-kbp window.

5.1 Retrotransposition in the human genome

Retrotransposition of processed mRNA transcripts into the genome results in functional genes, called intronless paralogs, or inactivated genes (pseudogenes). A paralog refers to a gene that appears in more than one copy in a given organism as a result of a duplication event. The existence of both intron-containing and intronless forms of genes encoding functionally similar or identical proteins has been previously described (*84*, *85*). Cataloging these evolutionary events on the genomic landscape is of value in understanding the functional consequences of such gene-duplication events in cellular biology. Identification of conserved intronless paralogs in the mouse or other mammalian genomes should provide the basis for capturing the evolutionary chronology of these transposition events and provide insights into gene loss and accretion in the mammalian radiation.

A set of proteins corresponding to all 901

Table 10. Features of the chromosomes. De novo/any refers to the union of de novo predictions that do not overlap Otto predictions and have at least one other type of supporting evidence; de novo/2x refers to the union of de novo predictions that do not overlap Otto predictions and have at least two types of evidence. Deserts are regions of sequence with no annotated genes.

	Sequence coverage (CS assembly)						Base composition			Gene prediction*					Gene density (genes/Mbp)						
Chr.	Size (Mbp)	No. of scaffolds	Largest scaffold (Mbp)	No. of scaffolds >500 kbp	Sequence covered by scaffolds >500 kbp	% of total sequence in scaffolds >500 kbp	% repeat	% GC	No of CpG islands	Otto	De novo/ any	De novo/ 2×	Total (Otto + de novo/ any)	Total (Otto + de novo/ any)	Sequence in deserts >500/ kbp	Sequence in deserts >1 Mbp	Otto	De novo/ any	De novo/ 2×	Otto + de novo/ any	Otto + de novo/ 2×
1	220	2,549	11	82	192	88	37	42	2,335	1,743	1,710	710	3,453	2,453	29	6	8	8	3	16	11
2	240	3,263	13	78	217	91	36	40	1,703	1,183	1,771	633	2,954	1,816	55	19	5	7	2	12	7
3	200	3,532	7	78	173	87	37	40	1,271	1,013	1,414	598	2,427	1,611	50	12	5	7	3	12	8
4	186	2,180	10	70	169	91	37	38	1,081	696	1,165	449	1,861	1,145	55	18	4	6	2	10	6
5	182	3,231	11	63	163	89	37	40	1,302	892	1,244	474	2,136	1,366	46	15	5	7	2	11	7
6	172	1,713	13	58	160	93	37	40	1,384	943	1,314	524	2,257	1,467	38	9	6	7	3	13	8
7	146	1,326	14	53	130	89	38	40	1,406	759	1,072	460	1,831	1,219	26	12	5	7	3	12	8
8	146	1,772	11	54	135	92	36	40	948	583	977	357	1,560	940	33	6	4	7	2	11	6
9	113	1,616	8	40	101	89	38	41	1,315	689	848	329	1,537	1,018	22	9	6	7	3	13	8
10	130	2,005	9	55	116	89	36	42	1,087	685	968	342	1,653	1,027	21	8	5	7	2	12	7
11	132	2,814	9	44	116	88	39	42	1,461	1,051	1,134	535	2,185	1,586	27	9	8	8	4	16	12
12	134	2,614	8	51	117	87	38	41	1,131	925	936	417	1,861	1,342	24	9	7	7	3	14	10
13	99	1,038	13	34	91	91	36	38	644	341	691	241	1,032	582	31	16	4	7	2	10	5
14	87	576	11	16	83	95	40	41	913	583	700	290	1,283	873	34	20	7	8	3	14	10
15	80	1,747	8	31	70	87	37	42	722	558	640	246	1,198	804	8	1	7	8	3	15	10
16	75	1,520	8	27	62	82	40	44	1,533	748	673	247	1,421	995	13	3	10	9	3	19	12
17	78	1,683	6	40	61	78	39	45	1,489	897	648	313	1,545	1,210	15	6	12	8	4	19	15
18	79	1,333	13	18	72	92	36	40	510	283	543	189	826	472	21	10	4	7	2	10	6
19	58	2,282	3	31	38	67	57	49	2,804	1,141	534	268	1,675	1,409	3	0	20	9	4	29	23
20	61	580	14	17	58	94	41	44	997	517	469	180	986	697	7	1	8	7	3	16	11
21	33	358	10	6	32	96	38	41	519	184	265	102	449	286	15	9	6	8	3	13	8
22	36	333	11	12	32	88	44	48	1,173	494	341	147	835	641	3	0	14	9	4	23	17
X	128	1,346	4	91	93	73	46	39	726	605	860	387	1,465	992	29	8	5	6	3	11	7
Y	19	638	2	10	12	65	50	39	65	55	155	49	210	104	4	2	3	8	2	11	5
U*	75	11,542	1						479	196	278	132	474	328							
Total	2907	53,591		1,059	2,490				28,519	17,764	21,350	8,619	39,114	26,383	606	208					
Avg.	116	2,144	9	44	104	87	40	41	1,160	714	812	333	1,526	1,047	25	9	7	7	3	14	9

*Chromosomal assignment unknown.

Otto-predicted, single-exon genes were subjected to BLAST analysis against the proteins encoded by the remaining multiexon predicted transcripts. Using homology criteria of 70% sequence identity over 90% of the length, we identified 298 instances of single- to multi-exon correspondence. Of these 298 sequences, 97 were represented in the GenBank data set of experimentally validated full-length genes at the stringency specified and were verified by manual inspection.

We believe that these 97 cases may represent intronless paralogs (see Web table 1 on *Science* Online at www.sciencemag.org/cgi/content/full/291/5507/1304/DC1) of known genes. Most of these are flanked by direct repeat sequences, although the precise nature of these repeats remains to be determined. All of the cases for which we have high confidence contain polyadenylated [poly(A)] tails characteristic of retrotransposition.

Recent publications describing the phenomenon of functional intronless paralogs speculate that retrotransposition may serve as a mechanism used to escape X-chromosomal inactivation (*84*, *86*). We do not find a bias toward X chromosome origination of these retrotransposed genes; rather, the results show a random chromosome distribution of both the intron-containing and corresponding intronless paralogs. We also have found several cases of retrotransposition from a single source chromosome to multiple target chromosomes. Interesting examples include the retrotransposition of a five exon–containing ribosomal protein L21 gene on chromosome 13 onto chromosomes 1, 3, 4, 7, 10, and 14, respectively. The size of the source genes can also show variability. The largest example is the 31-exon diacylglycerol kinase zeta gene on chromosome 11 that has an intronless paralog on chromosome 13. Regardless of route, retrotransposition with subsequent gene changes in coding or noncoding regions that lead to different functions or expression patterns, represents a key route to providing an enhanced functional repertoire in mammals (*87*).

Our preliminary set of retrotransposed intronless paralogs contains a clear overrepresentation of genes involved in translational processes (40% ribosomal proteins and 10% translation elongation factors) and nuclear regulation (HMG nonhistone proteins, 4%), as well as metabolic and regulatory enzymes. EST matches specific to a subset of intronless paralogs suggest expression of these intronless paralogs. Differences in the upstream regulatory sequences between the source genes and their intronless paralogs could account for differences in tissue-specific gene expression. Defining which, if any, of these processed genes are functionally expressed and translated will require further elucidation and experimental validation.

5.2 Pseudogenes

A pseudogene is a nonfunctional copy that is very similar to a normal gene but that has been altered slightly so that it is not expressed. We developed a method for the preliminary analysis of processed pseudogenes in the human genome as a starting point in elucidating the ongoing evolutionary forces

Table 11. Genome overview.

Size of the genome (including gaps)	2.91 Gbp
Size of the genome (excluding gaps)	2.66 Gbp
Longest contig	1.99 Mbp
Longest scaffold	14.4 Mbp
Percent of A+T in the genome	54
Percent of G+C in the genome	38
Percent of undetermined bases in the genome	9
Most GC-rich 50 kb	Chr. 2 (66%)
Least GC-rich 50 kb	Chr. X (25%)
Percent of genome classified as repeats	35
Number of annotated genes	26,383
Percent of annotated genes with unknown function	42
Number of genes (hypothetical and annotated)	39,114
Percent of hypothetical and annotated genes with unknown function	59
Gene with the most exons	Titin (234 exons)
Average gene size	27 kbp
Most gene-rich chromosome	Chr. 19 (23 genes/Mb)
Least gene-rich chromosomes	Chr. 13 (5 genes/Mb), Chr. Y (5 genes/Mb)
Total size of gene deserts (>500 kb with no annotated genes)	605 Mbp
Percent of base pairs spanned by genes	25.5 to 37.8*
Percent of base pairs spanned by exons	1.1 to 1.4*
Percent of base pairs spanned by introns	24.4 to 36.4*
Percent of base pairs in intergenic DNA	74.5 to 63.6*
Chromosome with highest proportion of DNA in annotated exons	Chr. 19 (9.33)
Chromosome with lowest proportion of DNA in annotated exons	Chr. Y (0.36)
Longest intergenic region (between annotated + hypothetical genes)	Chr. 13 (3,038,416 bp)
Rate of SNP variation	1/1250 bp

*In these ranges, the percentages correspond to the annotated gene set (26, 383 genes) and the hypothetical + annotated gene set (39,114 genes), respectively.

Table 12. Rate of recombination per physical distance (cM/Mb) across the genome. Genethon markers were placed on CSA-mapped assemblies, and then relative physical distances and rates were calculated in 3-Mb windows for each chromosome. NA, not applicable.

Chrom.	Male			Sex-average			Female		
	Max.	Avg.	Min.	Max.	Avg.	Min.	Max.	Avg.	Min.
1	2.60	1.12	0.23	2.81	1.42	0.52	3.39	1.76	0.68
2	2.23	0.78	0.33	2.65	1.12	0.54	3.17	1.40	0.61
3	2.55	0.86	0.23	2.40	1.07	0.42	2.71	1.30	0.33
4	1.66	0.67	0.15	2.06	1.04	0.60	2.50	1.40	0.77
5	2.00	0.67	0.18	1.87	1.08	0.42	2.26	1.43	0.62
6	1.97	0.71	0.28	2.57	1.12	0.37	3.47	1.67	0.64
7	2.34	1.16	0.48	1.67	1.17	0.47	2.27	1.21	0.34
8	1.83	0.73	0.14	2.40	1.05	0.46	3.44	1.36	0.43
9	2.01	0.99	0.53	1.95	1.32	0.77	2.63	1.66	0.82
10	3.73	1.03	0.22	3.05	1.29	0.66	2.84	1.51	0.76
11	1.43	0.72	0.31	2.13	0.99	0.47	3.10	1.32	0.49
12	4.12	0.76	0.26	3.35	1.16	0.49	2.93	1.55	0.59
13	1.60	0.75	0.01	1.87	0.95	0.17	2.49	1.19	0.32
14	3.15	0.98	0.18	2.65	1.30	0.62	3.14	1.63	0.75
15	2.28	0.94	0.34	2.31	1.22	0.42	2.53	1.56	0.54
16	1.83	1.00	0.47	2.70	1.55	0.63	4.99	2.32	1.12
17	3.87	0.87	0.00	3.54	1.35	0.54	4.19	1.83	0.94
18	3.12	1.37	0.86	3.75	1.66	0.43	4.35	2.24	0.72
19	3.02	0.97	0.10	2.57	1.41	0.49	2.89	1.75	0.87
20	3.64	0.89	0.00	2.79	1.50	0.83	3.31	2.15	1.34
21	3.23	1.26	0.69	2.37	1.62	1.08	2.58	1.90	1.18
22	1.25	1.10	0.84	1.88	1.41	1.08	3.73	2.08	0.93
X	NA	NA	NA	NA	NA	NA	3.12	1.64	0.72
Y	NA	NA	NA	NA	NA	NA	NA	NA	NA
Genome	4.12	0.88	0.00	3.75	1.22	0.17	4.99	1.55	0.32

that account for gene inactivation. The general structural characteristics of these processed pseudogenes include the complete lack of intervening sequences found in the functional counterparts, a poly(A) tract at the 3′ end, and direct repeats flanking the pseudogene sequence. Processed pseudogenes occur as a result of retrotransposition, whereas unprocessed pseudogenes arise from segmental genome duplication.

We searched the complete set of Otto-predicted transcripts against the genomic sequence by means of BLAST. Genomic regions corresponding to all Otto-predicted transcripts were excluded from this analysis. We identified 2909 regions matching with greater than 70% identity over at least 70% of the length of the transcripts that likely represent processed pseudogenes. This number is probably an underestimate because specific methods to search for pseudogenes were not used.

We looked for correlations between structural elements and the propensity for retrotransposition in the human genome. GC content and transcript length were compared between the genes with processed pseudogenes (1177 source genes) versus the remainder of the predicted gene set. Transcripts that give rise to processed pseudogenes have shorter average transcript length (1027 bp versus 1594 bp for the Otto set) as compared with genes for which no pseudogene was detected. The overall GC content did not show any significant difference, contrary to a recent report (*88*). There is a clear trend in gene families that are present as processed pseudogenes. These include ribosomal proteins (67%), lamin receptors (10%), translation elongation factor alpha (5%), and HMG–non-histone proteins (2%). The increased occurrence of retrotransposition (both intronless paralogs and processed pseudogenes) among genes involved in translation and nuclear regulation may reflect an increased transcriptional activity of these genes.

5.3 Gene duplication in the human genome

Building on a previously published procedure (*27*), we developed a graph-theoretic algorithm, called Lek, for grouping the predicted human protein set into protein families (*89*). The complete clusters that result from the Lek clustering provide one basis for comparing the role of whole-genome or chromosomal duplication in protein family expansion as opposed to other means, such as tandem duplication. Because each complete cluster represents a closed and certain island of homology, and because Lek is capable of simultaneously clustering protein complements of several organisms, the number of proteins contributed by each organism to a complete cluster can be predicted with confidence depending on the quality of the annotation of each genome. The variance of each organism's contribution to each cluster can then be calculated, allowing an assessment of the relative importance of large-scale duplication versus smaller-scale, organism-specific expansion and contraction of protein families, presumably as a result of natural selection operating on individual protein families within an organism. As can be seen in Fig. 12, the large variance in the relative numbers of human as compared with *D. melanogaster* and *Caenorhabditis elegans* proteins in complete clusters may be explained by multiple events of relative expansions in gene families in each of the three animal genomes. Such expansions would give rise to the distribution that shows a peak at 1:1 in the ratio for human-worm or human-fly clusters with the slope spread covering both human and fly/worm predominance, as we observed (Fig. 12). Furthermore, there are nearly as many clusters where worm and fly proteins predominate despite the larger numbers of proteins in the human. At face value, this analysis suggests that natural selection acting on individual protein families has been a major force driving the expansion of at least some elements of the human protein set. However, in our analysis, the difference between an ancient whole-genome duplication followed by loss, versus piecemeal duplication, cannot be easily distinguished. In order to differentiate these scenarios, more extended analyses were performed.

5.4 Large-scale duplications

Using two independent methods, we searched for large-scale duplications in the human genome. First, we describe a protein family–based method that identified highly conserved blocks of duplication. We then describe our comprehensive method for identifying all interchromosomal block duplications. The latter method identified a large number of duplicated chromosomal segments covering parts of all 24 chromosomes.

The first of the methods is based on the idea of searching for blocks of highly conserved homologous proteins that occur in more than one location on the genome. For this comparison, two genes were considered equivalent if their protein products were de-

Table 13. Characteristics of CpG islands identified in chromosome 22 (34-Mbp sequence length) and the whole genome (2.9-Gbp sequence length) by means of two different methods. Method 1 uses a CG likelihood ratio of ≥0.6. Method 2 uses a CG likelihood ratio of ≥0.8.

	Chromosome 22		Whole genome (CS assembly)	
	Method 1	Method 2	Method 1	Method 2
Number of CpG islands detected	5,211	522	195,706	26,876
Average length of island (bp)	390	535	395	497
Percent of sequence predicted as CpG	5.9	0.8	2.6	0.4
Percent of first exons that overlap a CpG island	44	25	42	22
Percent of first exons with first position of exon contained inside a CpG island	37	22	40	21
Average distance between first exon and closest CpG island (bp)	1,013	10,486	2,182	17,021
Expected distance between first exon and closest CpG island (bp)	3,262	32,567	7,164	55,811

Table 14. Distribution of repetitive DNA in the compartmentalized shotgun assembly sequence.

Repetitive elements	Megabases in assembled sequences	Percent of assembly	Previously predicted (%) (*83*)
Alu	288	9.9	10.0
Mammalian interspersed repeat (MIR)	66	2.3	1.7
Medium reiteration (MER)	50	1.7	1.6
Long terminal repeat (LTR)	155	5.3	5.6
Long interspersed nucleotide element (LINE)	466	16.1	16.7
Total	1025	35.3	35.6

termined to be in the same family and the same complete Lek cluster (essentially paralogous genes) (*89*). Initially, each chromosome was represented as a string of genes ordered by the start codons for predicted genes along the chromosome. We considered the two strands as a single string, because local inversions are relatively common events relative to large-scale duplications. Each gene was indexed according to the protein family and Lek complete cluster (*89*). All pairs of indexed gene strings were then aligned in both the forward and reverse directions with the Smith-Waterman algorithm (*90*). A match between two proteins of the same Lek complete cluster was given a score of 10 and a mismatch −10, with gap open and extend penalties of −4 and −1. With these parameters, 19 conserved interchromosomal blocks of duplication were observed, all of which were also detected and expanded by the comprehensive method described below. The detection of only a relatively small number of block duplications was a consequence of using an intrinsically conservative method grounded in the conservative constraints of the complete Lek clusters.

In the second, more comprehensive approach, we aligned all chromosomes directly with one another using an algorithm based on the MUMmer system (*91*). This alignment method uses a suffix tree data structure and a linear-time algorithm to align long sequences very rapidly; for example, two chromosomes of 100 Mbp can be aligned in less than 20 min (on a Compaq Alpha computer) with 4 gigabytes of memory. This procedure was used recently to identify numerous large-scale segmental duplications among the five chromosomes of *A. thaliana* (*92*); in that organism, the method revealed that 60% of the genome (66 Mbp) is covered by 24 very large duplicated segments. For *Arabidopsis*, a DNA-based alignment was sufficient to reveal the segmental duplications between chromosomes; in the human genome, DNA alignments at the whole-chromosome level are insufficiently sensitive. Therefore, a modified procedure was developed and applied, as follows. First, all 26,588 proteins (9,675,713 million amino acids) were concatenated end-to-end in order as they occur along each of the 24 chromosomes, irrespective of strand location. The concatenated protein set was then aligned against each chromosome by the MUMmer algorithm. The resulting matches were clustered to extract all sets of three or more protein matches that occur in close proximity on two different chromosomes (*93*); these represent the candidate segmental duplications. A series of filters were developed and applied to remove likely false-positives from this set; for example, small blocks that were spread across many proteins were removed. To refine the filtering methods, a shuffled protein set was first created by taking the 26,588 proteins, randomizing their order, and then partitioning them into 24 shuffled chromosomes, each containing the same number of proteins as the true genome. This shuffled protein set has the identical composition to the real genome; in particular, every protein and every domain appears the same number of times. The complete algorithm was then applied to both the real and the shuffled data, with the results on the shuffled data being used to estimate the false-positive rate. The algorithm after filtering yielded 10,310 gene pairs in 1077 duplicated blocks containing 3522 distinct genes; tandemly duplicated expansions in many of the blocks explain the excess of gene pairs to distinct genes. In the shuffled data, by contrast, only 370 gene pairs were found, giving a false-positive estimate of 3.6%. The most likely explanation for the 1077 block duplications is ancient segmental duplications. In many cases, the order of the proteins has been shuffled, although proximity is preserved. Out of the 1077 blocks, 159 contain only three genes, 137 contain four genes, and 781 contain five or more genes.

To illustrate the extent of the detected duplications, Fig. 13 shows all 1077 block duplications indexed to each chromosome in 24 panels in which only duplications mapped to the indexed chromosome are displayed. The figure makes it clear that the duplications are ubiquitous in the genome. One feature that it displays is many relatively small chromosomal stretches, with one-to-many duplication relationships that are graphically striking. One such example captured by the analysis is the well-documented olfactory receptor (OR) family, which is scattered in blocks throughout the genome and which has been analyzed for genome-deployment reconstructions at several evolutionary stages (*94*). The figure also illustrates that some chromosomes, such as chromosome 2, contain many more detected large-scale duplications than others. Indeed, one of the largest duplicated segments is a large block of 33 proteins on chromosome 2, spread among eight smaller blocks in 2p, that aligns to a paralogous set on chromosome 14, with one rearrangement (see chromosomes 2 and 14 panels in Fig. 13). The proteins are not contiguous but span a region containing 97 proteins on chromosome 2 and 332 proteins on chromosome 14. The likelihood of observing this many duplicated proteins by chance, even over a span of this length, is 2.3×10^{-68} (*93*). This duplicated set spans 20 Mbp on chromosome 2 and 63 Mbp on chromosome 14, over 70% of the latter chromosome. Chromosome 2 also contains a block duplication that is nearly as large, which is shared by chromosome arm 2q and chromosome 12. This duplication incorporates two of the four known Hox gene clusters, but considerably expands the extent of the duplications proximally and distally on the pair of chromosome arms. This breadth of duplication is also seen on the two chromosomes carrying the other two Hox clusters.

An additional large duplication, between chromosomes 18 and 20, serves as a good example to illustrate some of the features common to many of the other observed large duplications (Fig. 13, inset). This duplication contains 64 detected ordered intrachromosomal pairs of homologous genes. After discounting a 40-Mb stretch of chromosome 18 free of matches to chromosome 20, which is likely to represent a large insert (between the gene assignments "Krup rel" and "collagen rel" on chromosome 18 in Fig. 13), the full duplication segment covers 36 Mb on chromosome 18 and 28 Mb on chromosome 20.

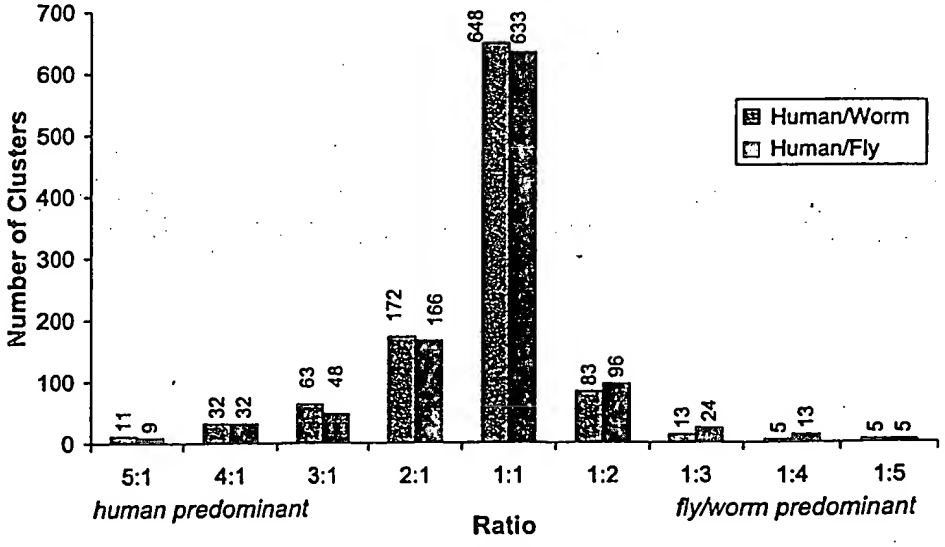


Fig. 12. Gene duplication in complete protein clusters. The predicted protein sets of human, worm, and fly were subjected to Lek clustering (*27*). The numbers of clusters with varying ratios (whole number) of human versus worm and human versus fly proteins per cluster were plotted.

By this measure, the duplication segment spans nearly half of each chromosome's net length. The most likely scenario is that the whole span of this region was duplicated as a single very large block, followed by shuffling owing to smaller scale rearrangements. As such, at least four subsequent rearrangements would need to be invoked to explain the relative insertions and inversions seen in the duplicated segment interval. The 64 protein pairs in this alignment occur among 217 protein assignments on chromosome 18, and among 322 protein assignments on chromosome 20, for a density of involved proteins of 20 to 30%. This is consistent with an ancient large-scale duplication followed by subsequent gene loss on one or both chromosomes. Loss of just one member of a gene pair subsequent to the duplication would result in a failure to score a gene pair in the block; less than 50% gene loss on the chromosomes would lead to the duplication density observed here. As an independent verification of the significance of the alignments detected, it can be seen that a substantial number of the pairs of aligning proteins in this duplication, including some of those annotated (Fig. 13), are those populating small Lek complete clusters (see above). This indicates that they are members of very small families of paralogs; their relative scarcity within the genome validates the uniqueness and robust nature of their alignments.

Two additional qualitative features were observed among many of the large-scale duplications. First, several proteins with disease associations, with OMIM (Online Mendelian Inheritance in Man) assignments, are members of duplicated segments (see web table 2 on *Science* Online at www.sciencemag.org/cgi/content/full/291/5507/1304/DC1). We have also observed a few instances where paralogs on both duplicated segments are associated with similar disease conditions. Notable among these genes are proteins involved in hemostasis (coagulation factors) that are associated with bleeding disorders, transcriptional regulators like the homeobox proteins associated with developmental disorders, and potassium channels associated with cardiovascular conduction abnormalities. For each of these disease genes, closer study of the paralogous genes in the duplicated segment may reveal new insights into disease causation, with further investigation needed to determine whether they might be involved in the same or similar genetic diseases. Second, although there is a conserved number of proteins and coding exons predicted for specific large duplicated spans within the chromosome 18 to 20 alignment, the genomic DNA of chromosome 18 in these specific spans is in some cases more than 10-fold longer than the corresponding chromosome 20 DNA. This selective accretion of noncoding DNA (or conversely, loss of noncoding DNA) on one of a pair of duplicated chromosome regions was observed in many compared regions. Hypotheses to explain which mechanisms foster these processes must be tested.

Evaluation of the alignment results gives some perspective on dating of the duplications. As noted above, large-scale ancient segmental duplication in fact best explains many of the blocks detected by this genome-wide analysis. The regions of human chromosomes involved in the large-scale duplications expanded upon above (chromosomes 2 to 14, 2 to 12, and 18 to 20) are each syntenic to a distinct mouse chromosomal region. The corresponding mouse chromosomal regions are much more similar in sequence conservation, and even in order, to their human synteny partners than the human duplication regions are to each other. Further, the corresponding mouse chromosomal regions each bear a significant proportion of genes orthologous to the human genes on which the human duplication assignments were made. On the basis of these factors, the corresponding mouse chromosomal spans, at coarse resolution, appear to be products of the same large-scale duplications observed in humans. Although further detailed analysis must be carried out once a more complete genome is assembled for mouse, the underlying large duplications appear to predate the two species' divergence. This dates the duplications, at the latest, before divergence of the primate and rodent lineages. This date can be further refined upon examination of the synteny between human chromosomes and those of chicken, pufferfish (*Fugu rubripes*), or zebrafish (*95*). The only substantial syntenic stretches mapped in these species corresponding to both pairs of human duplications are restricted to the Hox cluster regions. When the synteny of these regions (or others) to human chromosomes is extended with further mapping, the ages of the nearly chromosome-length duplications seen in humans are likely to be dated to the root of vertebrate divergence.

The MUMmer-based results demonstrate large block duplications that range in size from a few genes to segments covering most of a chromosome. The extent of segmental duplications raises the question of whether an ancient whole-genome duplication event is the underlying explanation for the numerous duplicated regions (*96*). The duplications have undergone many deletions and subsequent rearrangements; these events make it difficult to distinguish between a whole-genome duplication and multiple smaller events. Further analysis, focused especially on comparing the estimated ages of all the block duplications, derived partially from interspecies genome comparisons, will be necessary to determine which of these two hypotheses is more likely. Comparisons of genomes of different vertebrates, and even cross-phyla genome comparisons, will allow for the deconvolution of duplications to eventually reveal the stagewise history of our genome, and with it a history of the emergence of many of the key functions that distinguish us from other living things.

6 A Genome-Wide Examination of Sequence Variations

Summary. Computational methods were used to identify single-nucleotide polymorphisms (SNPs) by comparison of the Celera sequence to other SNP resources. The SNP rate between two chromosomes was ~1 per 1200 to 1500 bp. SNPs are distributed nonrandomly throughout the genome. Only a very small proportion of all SNPs (<1%) potentially impact protein function based on the functional analysis of SNPs that affect the predicted coding regions. This results in an estimate that only thousands, not millions, of genetic variations may contribute to the structural diversity of human proteins.

Having a complete genome sequence enables researchers to achieve a dramatic acceleration in the rate of gene discovery, but only through analysis of sequence variation in DNA can we discover the genetic basis for variation in health among human beings. Whole-genome shotgun sequencing is a particularly effective method for detecting sequence variation in tandem with whole-genome assembly. In addition, we compared the distribution and attributes of SNPs ascertained by three other methods: (i) alignment of the Celera consensus sequence to the PFP assembly, (ii) overlap of high-quality reads of genomic sequence (referred to as "Kwok"; 1,120,195 SNPs) (*97*), and (iii) reduced representation shotgun sequencing (referred to as "TSC"; 632,640 SNPs) (*98*). These data were consistent in showing an overall nucleotide diversity of $\sim 8 \times 10^{-4}$, marked heterogeneity across the genome in SNP density, and an overwhelming preponderance of noncoding variation that produces no change in expressed proteins.

6.1 SNPs found by aligning the Celera consensus to the PFP assembly

Ideally, methods of SNP discovery make full use of sequence depth and quality at every site, and quantitatively control the rate of false-positive and false-negative calls with an explicit sampling model (*99*). Comparison of consensus sequences in the absence of these details necessitated a more ad hoc approach (quality scores could not readily be obtained for the PFP assembly). First, all sequence differences between the two consensus sequences were identified; these were then filtered to reduce the contribution of sequencing errors and misassembly. As a measure of the effectiveness of the filtering step, we monitored the ratio of transition and transversion substitutions, because a 2:1 ratio has been well documented as typical in mammalian evolution (*100*) and in human SNPs

(*101*, *102*). The filtering steps consisted of removing variants where the quality score in the Celera consensus was less than 30 and where the density of variants was greater than 5 in 400 bp. These filters resulted in shifting the transition-to-transversion ratio from 1.57:1 to 1.89:1. When applied to 2.3 Gbp of alignments between the Celera and PFP consensus sequences, these filters resulted in identification of 2,104,820 putative SNPs from a total of 2,778,474 substitution differences. Overlaps between this set of SNPs and those found by other methods are described below.

6.2 Comparisons to public SNP databases

Additional SNPs, including 2,536,021 from dbSNP (www.ncbi.nlm.nih.gov/SNP) and 13,150 from HGMD (Human Gene Mutation Database, from the University of Wales, UK), were mapped on the Celera consensus sequence by a sequence similarity search with the program PowerBlast (*103*). The two largest data sets in dbSNP are the Kwok and TSC sets, with 47% and 25% of the dbSNP records. Low-quality alignments with partial coverage of the dbSNP sequence and alignments that had less than 98% sequence identity between the Celera sequence and the dbSNP flanking sequence were eliminated. dbSNP sequences mapping to multiple locations on the Celera genome were discarded. A total of 2,336,935 dbSNP variants were mapped to 1,223,038 unique locations on the Celera sequence, implying considerable redundancy in dbSNP. SNPs in the TSC set mapped to 585,811 unique genomic locations, and SNPs in the Kwok set mapped to 438,032 unique locations. The combined unique SNPs counts used in this analysis, including Celera-PFP, TSC, and Kwok, is 2,737,668. Table 15 shows that a substantial fraction of SNPs identified by one of these methods was also found by another method. The very high overlap (36.2%) between the Kwok and Celera-PFP SNPs may be due in part to the use by Kwok of sequences that went into the PFP assembly. The unusually low overlap (16.4%) between the Kwok and TSC sets is due to their being the smallest two sets. In addition, 24.5% of the Celera-PFP SNPs overlap with SNPs derived from the Celera genome sequences (*46*). SNP validation in population samples is an expensive and laborious process, so confirmation on multiple data sets may provide an efficient initial validation "in silico" (by computational analysis).

One means of assessing whether the three sets of SNPs provide the same picture of human variation is to tally the frequencies of the six possible base changes in each set of SNPs (Table 16). Previous measures of nucleotide diversity were mostly derived from small-scale analysis on candidate genes (*101*), and our analysis with all three data sets validates the previous observations at the whole-genome scale. There is remarkable homogeneity between the SNPs found in the Kwok set, the TSC set, and in our whole-genome shotgun (*46*) in this substitution pattern. Compared with the rest of the data sets, Celera-PFP deviates slightly from the 2:1 transition-to-transversion ratio observed in the other SNP sets. This result is not unexpected, because some fraction of the computationally identified SNPs in the Celera-PFP comparison may in fact be sequence errors. A 2:1 transition:transversion ratio for the bona fide SNPs would be obtained if one assumed that 15% of the sequence differences in the Celera-PFP set were a result of (presumably random) sequence errors.

6.3 Estimation of nucleotide diversity from ascertained SNPs

The number of SNPs identified varied widely across chromosomes. In order to normalize these values to the chromosome size and sequence coverage, we used π, the standard statistic for nucleotide diversity (*104*). Nucleotide diversity is a measure of per-site heterozygosity, quantifying the probability that a pair of chromosomes drawn from the population will differ at a nucleotide site. In order to calculate nucleotide diversity for each chromosome, we need to know the number of nucleotide sites that were surveyed for variation, and in methods like reduced respresentation sequencing, we need to know the sequence quality and the depth of coverage at each site. These data are not readily available, so we could not estimate nucleotide diversity from the TSC effort. Estimation of nucleotide diversity from high-quality sequence overlaps should be possible, but again, more information is needed on the details of all the alignments.

Estimation of nucleotide diversity from a shotgun assembly entails calculating for each column of the multialignment, the probability that two or more distinct alleles are present, and the probability of detecting a SNP if in fact the alleles have different sequence (i.e., the probability of correct sequence calls). The greater the depth of coverage and the higher the sequence quality, the higher is the chance of successfully detecting a SNP (*105*). Even after correcting for variation in coverage, the nucleotide diversity appeared to vary across autosomes. The significance of this heterogeneity was tested by analysis of variance, with estimates of π for 100-kbp windows to estimate variability within chromosomes (for the Celera-PFP comparison, $F = 29.73$, $P < 0.0001$).

Average diversity for the autosomes estimated from the Celera-PFP comparison was 8.94×10^{-4}. Nucleotide diversity on the X chromosome was 6.54×10^{-4}. The X is expected to be less variable than autosomes, because for every four copies of autosomes in the population, there are only three X chromosomes, and this smaller effective population size means that random drift will more rapidly remove variation from the X (*106*).

Having ascertained nucleotide variation genome-wide, it appears that previous estimates of nucleotide diversity in humans based on samples of genes were reasonably accurate (*101*, *102*, *106*, *107*). Genome-wide, our estimate of nucleotide diversity was 8.98×10^{-4} for the Celera-PFP alignment, and a published estimate averaged over 10 densely resequenced human genes was 8.00×10^{-4} (*108*).

6.4 Variation in nucleotide diversity across the human genome

Such an apparently high degree of variability among chromosomes in SNP density raises the question of whether there is heterogeneity at a finer scale within chromo-

Table 15. Overlap of SNPs from genome-wide SNP databases. Table entries are SNP counts for each pair of data sets. Numbers in parentheses are the fraction of overlap, calculated as the count of overlapping SNPs divided by the number of SNPs in the smaller of the two databases compared. Total SNP counts for the databases are: Celera-PFP, 2,104,820; TSC, 585,811; and Kwok 438,032. Only unique SNPs in the TSC and Kwok data sets were included.

	TSC	Kwok
Celera-PFP	188,694 (0.322)	158,532 (0.362)
		72,024 (0.164)

Table 16. Summary of nucleotide changes in different SNP data sets.

SNP data set	A/G (%)	C/T (%)	A/C (%)	A/T (%)	C/G (%)	T/G (%)	Transition: transversion
Celera-PFP	30.7	30.7	10.3	8.6	9.2	10.3	1.59:1
Kwok*	33.7	33.8	8.5	7.0	8.6	8.4	2.07:1
TSC†	33.3	33.4	8.8	7.3	8.6	8.6	1.99:1

*November 2000 release of the NCBI database dbSNP (www.ncbi.nlm.nih.gov/SNP/) with the method defined as Overlap SnpDetectionWithPolyBayes. The submitter of the data is Pui-Yan Kwok from Washington University. †November 2000 release of NCBI dbSNP (www.ncbi.nlm.nih.gov/SNP/) with the methods defined as TSC-Sanger, TSC-WICGR, and TSC-WUGSC. The submitter of the data is Lincoln Stein from Cold Spring Harbor Laboratory.

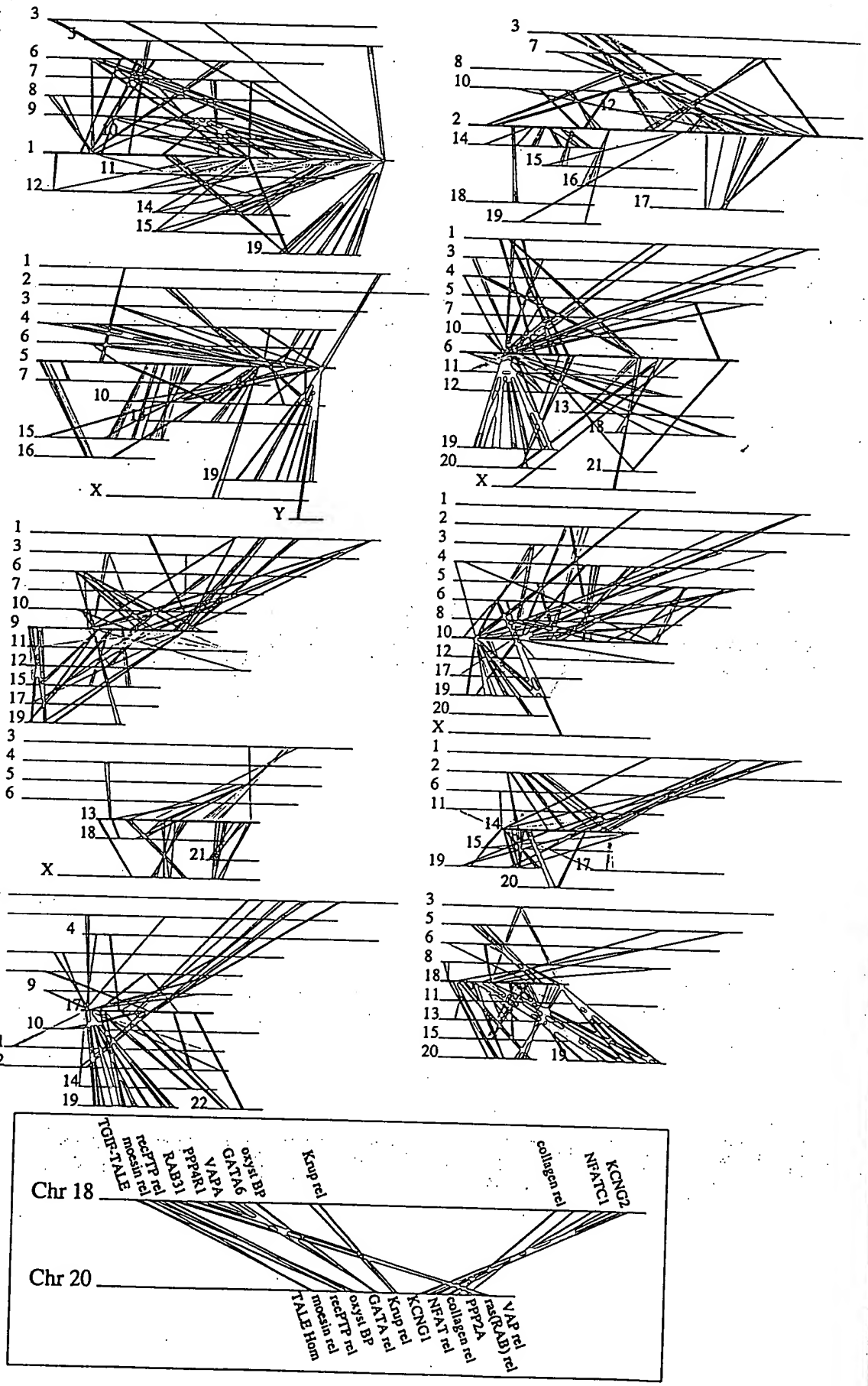
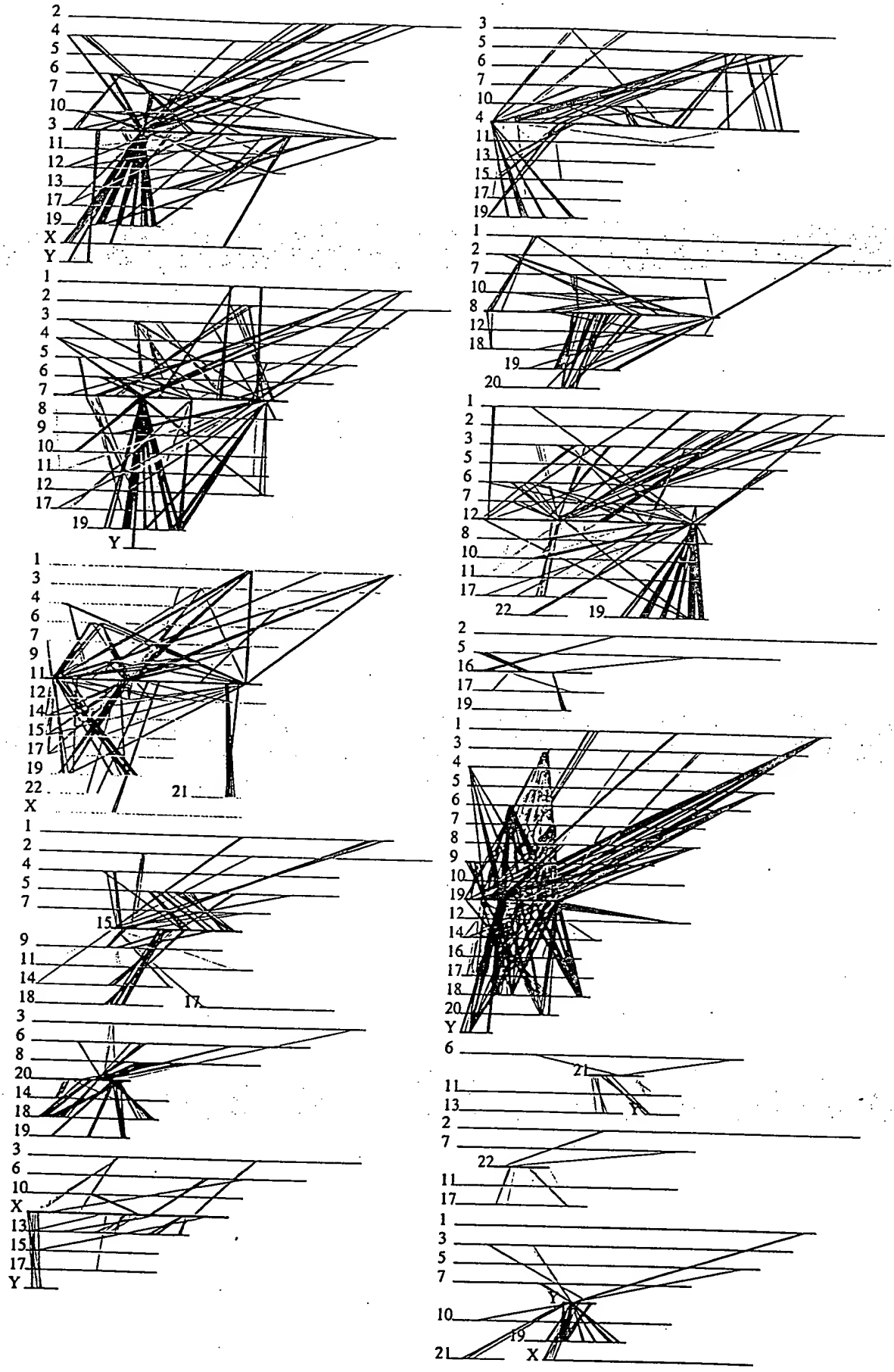


Fig. 13. Segmental duplications between chromosomes in the human genome. The 24 panels show the 1077 duplicated blocks of genes, containing 10,310 pairs of genes in total. Each line represents a pair of homologous genes belonging to a block; all blocks contain at least three genes on each of the chromosomes where they appear. Each panel shows all the duplications between a single chromosome and other chromosomes with shared blocks. The chromosome at the center of each panel is shown as a thick red line for emphasis. Other chromosomes are displayed from top to bottom within each panel ordered by chromosome number. The inset (bottom, center right) shows a close-up of one duplication between chromosomes 18 and 20, expanded to display the gene names of 12 of the 64 gene pairs shown.



somes, and whether this heterogeneity is greater than expected by chance. If SNPs occur by random and independent mutations, then it would seem that there ought to be a Poisson distribution of numbers of SNPs in fragments of arbitrary constant size. The observed dispersion in the distribution of SNPs in 100-kbp fragments was far greater than predicted from a Poisson distribution (Fig. 14). However, this simplistic model ignores the different recombination rates and population histories that exist in different regions of the genome. Population genetics theory holds that we can account for this variation with a mathematical formulation called the neutral coalescent (*109*). Applying well-tested algorithms for simulating the neutral coalescent with recombination (*110*), and using an effective population size of 10,000 and a per-base recombination rate equal to the mutation rate (*111*), we generated a distribution of numbers of SNPs by this model as well (*112*). The observed distribution of SNPs has a much larger variance than either the Poisson model or the coalescent model, and the difference is highly significant. This implies that there is significant variability across the genome in SNP density, an observation that begs an explanation.

Several attributes of the DNA sequence may affect the local density of SNPs, including the rate at which DNA polymerase makes errors and the efficacy of mismatch repair. One key factor that is likely to be associated with SNP density is the G+C content, in part because methylated cytosines in CpG dinucleotides tend to undergo deamination to form thymine, accounting for a nearly 10-fold increase in the mutation rate of CpGs over other dinucleotides. We tallied the GC content and nucleotide diversities in 100-kbp windows across the entire genome and found that the correlation between them was positive (r = 0.21) and highly significant ($P < 0.0001$), but G+C content accounted for only a small part of the variation.

6.5 SNPs by genomic class

To test homogeneity of SNP densities across functional classes, we partitioned sites into intergenic (defined as >5 kbp from any predicted transcription unit), 5′-UTR, exonic (missense and silent), intronic, and 3′-UTR for 10,239 known genes, derived from the NCBI RefSeq database and all human genes predicted from the Celera Otto annotation. In coding regions, SNPs were categorized as either silent, for those that do not change amino acid sequence, or missense, for those that change the protein product. The ratio of missense to silent coding SNPs in Celera-PFP, TSC, and Kwok sets (1.12, 0.91, and 0.78, respectively) shows a markedly reduced frequency of missense variants compared with the neutral expectation, consistent with the elimination by natural selection of a fraction of the deleterious amino acid changes (*112*). These ratios are comparable to the missense-to-silent ratios of 0.88 and 1.17 found by Cargill *et al.* (*101*) and by Halushka *et al.* (*102*). Similar results were observed in SNPs derived from Celera shotgun sequences (*46*).

It is striking how small is the fraction of SNPs that lead to potentially dysfunctional alterations in proteins. In the 10,239 RefSeq genes, missense SNPs were only about 0.12, 0.14, and 0.17% of the total SNP counts in Celera-PFP, TSC, and Kwok SNPs, respectively. Nonconservative protein changes constitute an even smaller fraction of missense SNPs (47, 41, and 40% in Celera-PFP, Kwok, and TSC). Intergenic regions have been virtually unstudied (*113*), and we note that 75% of the SNPs we identified were intergenic (Table 17). The SNP rate was highest in introns and lowest in exons. The SNP rate was lower in intergenic regions than in introns, providing one of the first discriminators between these two classes of DNA. These SNP rates were confirmed in the Celera SNPs, which also exhibited a lower rate in exons than in introns, and in extragenic regions than in introns (*46*). Many of these intergenic SNPs will provide valuable information in the form of markers for linkage and association studies, and some fraction is likely to have a regulatory function as well.

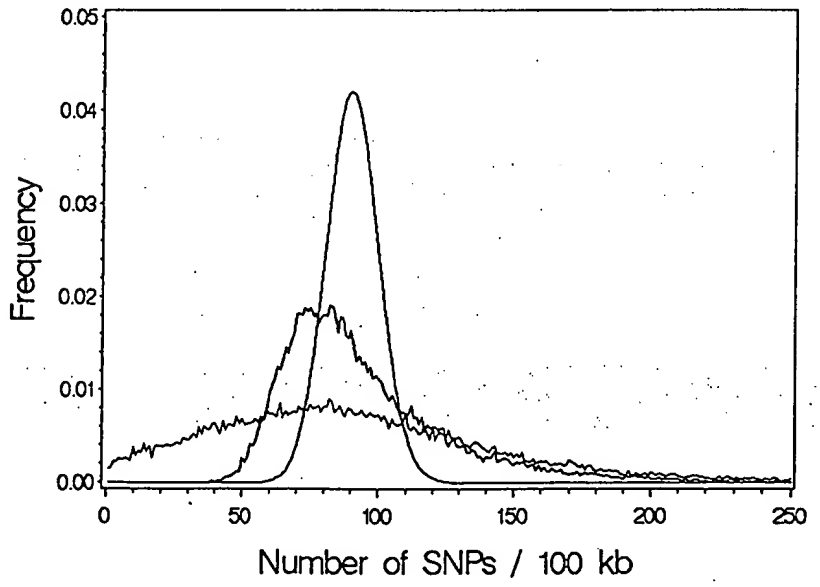


Fig. 14. SNP density in each 100-kbp interval as determined with Celera-PFP SNPs. The color codes are as follows: black, Celera-PFP SNP density; blue, coalescent model; and red, Poisson distribution. The figure shows that the distribution of SNPs along the genome is nonrandom and is not entirely accounted for by a coalescent model of regional history.

7 An Overview of the Predicted Protein-Coding Genes in the Human Genome

Summary. This section provides an initial computational analysis of the predicted protein set with the aim of cataloging prominent differences and similarities when the human genome is compared with other fully sequenced eukaryotic genomes. Over 40% of the predicted protein set in humans cannot be ascribed a molecular function by methods that assign proteins to known families. A protein domain–based analysis provides a detailed catalog of the prominent differences in the human genome when compared with the fly and worm genomes. Prominent among these are domain expansions in proteins involved in developmental regulation and in cellular processes such as neuronal function, hemostasis, acquired immune response, and cytoskeletal complexity. The final enumeration of protein families and details of protein structure will rely on additional experimental work and comprehensive manual curation.

A preliminary analysis of the predicted human protein-coding genes was conducted. Two methods were used to analyze and classify the molecular functions of 26,588 predicted proteins that represent 26,383 gene predictions with at least two lines of evidence as described above. The first method was based on an analysis at the level of protein families, with both the publicly available Pfam database (*114*, *115*) and Celera's Panther Classification (CPC) (Fig. 15) (*116*). The second method was based on an analysis at the level of protein domains, with both the Pfam and SMART databases (*115*, *117*).

The results presented here are preli-
nary and are subject to several limit-

Both the gene predictions and functional assignments have been made by using computational tools, although the statistical models in Panther, Pfam, and SMART have been built, annotated, and reviewed by expert biologists. In the set of computationally predicted genes, we expect both false-positive predictions (some of these may in fact be inactive pseudogenes) and false-negative predictions (some human genes will not be computationally predicted). We also expect errors in delimiting the boundaries of exons and genes. Similarly, in the automatic functional assignments, we also expect both false-positive and false-negative predictions. The functional assignment protocol focuses on protein families that tend to be found across several organisms, or on families of known human genes. Therefore, we do not assign a function to many genes that are not in large families, even if the function is known. Unless otherwise specified, all enumeration of the genes in any given family or functional category was taken from the set of 26,588 predicted proteins, which were assigned functions by using statistical score cutoffs defined for models in Panther, Pfam, and SMART.

For this initial examination of the predicted human protein set, three broad questions were asked: (i) What are the likely molecular functions of the predicted gene products, and how are these proteins categorized with current classification methods? (ii) What are the core functions that appear to be common across the animals? (iii) How does the human protein complement differ from that of other sequenced eukaryotes?

7.1 Molecular functions of predicted human proteins

Figure 15 shows an overview of the putative molecular functions of the predicted 26,588 human proteins that have at least two lines of supporting evidence. About 41% (12,809) of the gene products could not be classified from this initial analysis and are termed proteins with unknown functions. Because our automatic classification methods treat only relatively large protein families, there are a number of "unclassified" sequences that do, in fact, have a known or predicted function. For the 60% of the protein set that have automatic functional predictions, the specific protein functions have been placed into broad classes. We focus here on molecular function (rather than higher order cellular processes) in order to classify as many proteins as possible. These functional predictions are based on similarity to sequences of known function.

In our analysis of the 12,731 additional low-confidence predicted genes (those with only one piece of supporting evidence), only 636 (5%) of these additional putative genes were assigned molecular functions by the automated methods. One-third of these 636 predicted genes represented endogenous retroviral proteins, further suggesting that the majority of these unknown-function genes are not real genes. Given that most of these additional 12,095 genes appear to be unique among the genomes sequenced to date, many may simply represent false-positive gene predictions.

The most common molecular functions are the transcription factors and those involved in nucleic acid metabolism (nucleic acid enzyme). Other functions that are highly represented in the human genome are the receptors, kinases, and hydrolases. Not surprisingly, most of the hydrolases are proteases. There are also many proteins that are members of proto-oncogene families, as well as families of "select regulatory molecules": (i) proteins involved in specific steps of signal transduction such as heterotrimeric GTP-binding proteins (G proteins) and cell cycle regulators, and (ii) proteins that modulate the activity of kinases, G proteins, and phosphatases.

Table 17. Distribution of SNPs in classes of genomic regions.

Genomic region class	Size of region examined (Mb)	Celera-PFP SNP density (SNP/Mb)
Intergenic	2185	707
Gene (intron + exon)	646	917
Intron	615	921
First intron	164	808
Exon	31	529
First exon	10	592

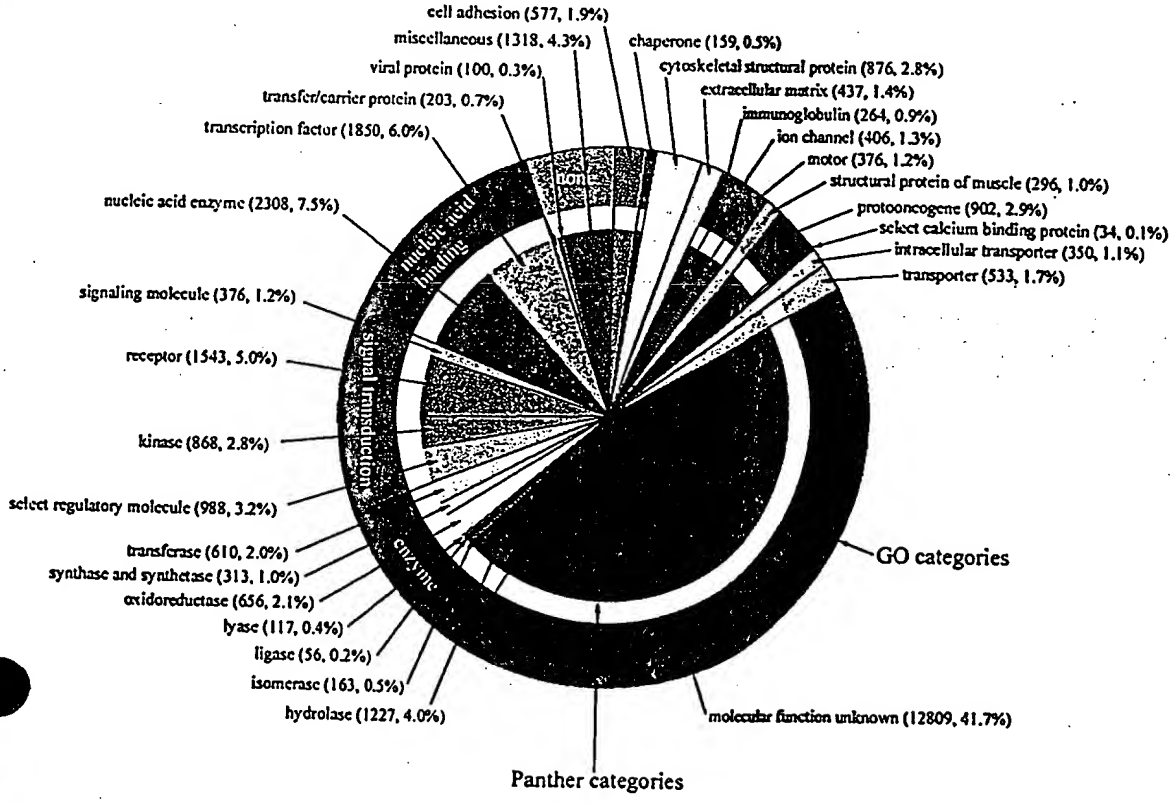


Fig. 15. Distribution of the molecular functions of 26,383 human genes. Each slice lists the numbers and percentages (in parentheses) of human gene functions assigned to a given category of molecular function. The outer circle shows the assignment to molecular function categories in the Gene Ontology (GO) (*179*), and the inner circle shows the assignment to Celera's Panther molecular function categories (*116*).

7.2 Evolutionary conservation of core processes

Because of the various "model organism" genome-sequencing projects that have already been completed, reasonable comparative information is available for beginning the analysis of the evolution of the human genome. The genomes of *S. cerevisiae* ("bakers' yeast") (*118*) and two diverse invertebrates, *C. elegans* (a nematode worm) (*119*) and *D. melanogaster* (fly) (*26*), as well as the first plant genome, *A. thaliana*, recently completed (*92*), provide a diverse background for genome comparisons.

We enumerated the "strict orthologs" conserved between human and fly, and between human and worm (Fig. 16) to address the question, What are the core functions that appear to be common across the animals? The concept of orthology is important because if two genes are orthologs, they can be traced by descent to the common ancestor of the two organisms (an "evolutionarily conserved protein set"), and therefore are likely to perform similar conserved functions in the different organisms. It is critical in this analysis to separate orthologs (a gene that appears in two organisms by descent from a common ancestor) from paralogs (a gene that appears in more than one copy in a given organism by a duplication event) because paralogs may subsequently diverge in function. Following the yeast-worm ortholog comparison in (*120*), we identified two different cases for each pairwise comparison (human-fly and human-worm). The first case was a pair of genes, one from each organism, for which there was no other close homolog in either organism. These are straightforwardly identified as orthologous, because there are no additional members of the families that complicate separating orthologs from paralogs. The second case is a family of genes with more than one member in either or both of the organisms being compared. Chervitz *et al.* (*120*) deal with this case by analyzing a phylogenetic tree that described the relationships between all of the sequences in both organisms, and then looked for pairs of genes that were nearest neighbors in the tree. If the nearest-neighbor pairs were from different organisms, those genes were presumed to be orthologs. We note that these nearest neighbors can often be confidently identified from pairwise sequence comparison without having to examine a phylogenetic tree (see legend to Fig. 16). If the nearest neighbors are not from different organisms, there has been a paralogous expansion in one or both organisms after the speciation event (and/or a gene loss by one organism). When this one-to-one correspondence is lost, defining an ortholog becomes ambiguous. For our initial computational overview of the predicted human protein set, we could not answer this question for every predicted protein. Therefore, we consider only "strict orthologs," i.e., the proteins with unambiguous one-to-one relationships (Fig. 16). By these criteria, there are 2758 strict human-fly orthologs, 2031 human-worm (1523 in common between these sets). We define the evolutionarily conserved set as those 1523 human proteins that have strict orthologs in both *D. melanogaster* and *C. elegans*.

The distribution of the functions of the conserved protein set is shown in Fig. 16. Comparison with Fig. 15 shows that, not surprisingly, the set of conserved proteins is not distributed among molecular functions in the same way as the whole human protein set. Compared with the whole human set (Fig. 15), there are several categories that are overrepresented in the conserved set by a factor of ~2 or more. The first category is nucleic acid enzymes, primarily the transcriptional machinery (notably DNA/RNA methyltransferases, DNA/RNA polymerases, helicases, DNA ligases, DNA- and RNA-processing factors, nucleases, and ribosomal proteins). The basic transcriptional and translational machinery is well known to have been conserved over evolution, from bacteria through to the most complex eukaryotes. Many ribonucleoproteins involved in RNA splicing also appear to be conserved among the animals. Other enzyme types are also overrepresented (transferases, oxidoreductases, ligases, lyases, and isomerases). Many of these en-

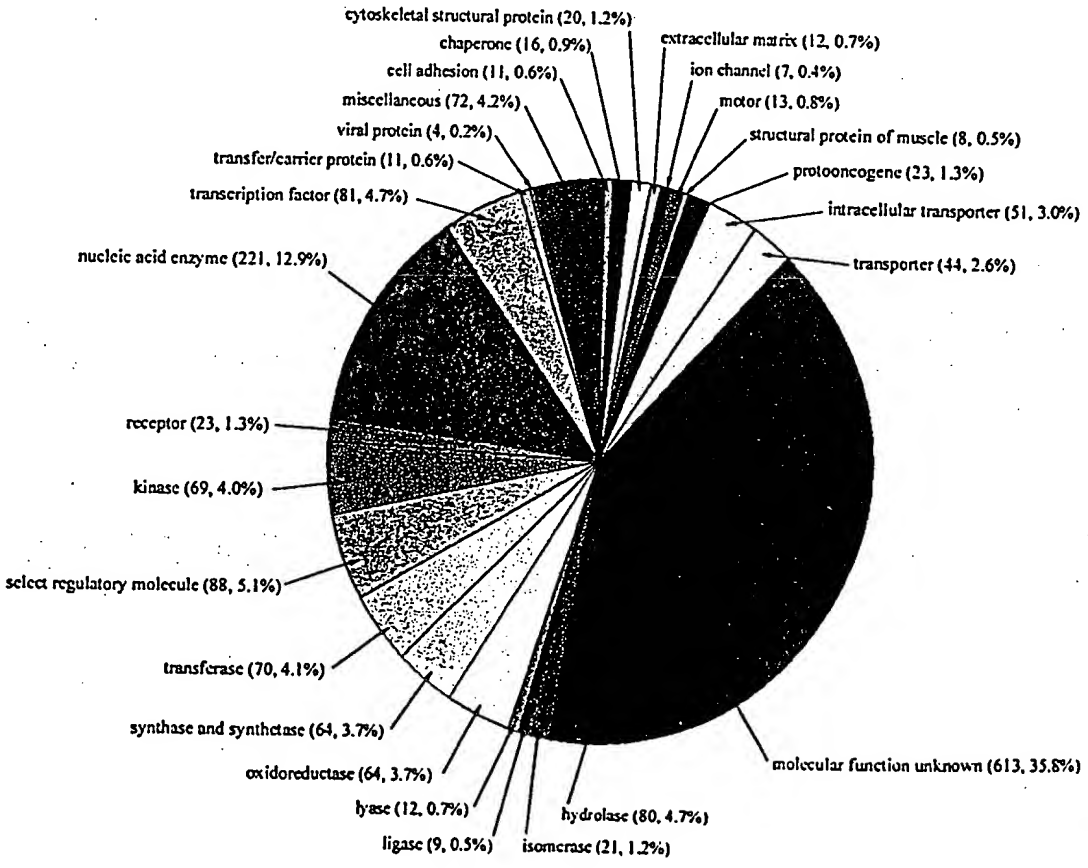


Fig. 16. Functions of putative orthologs across vertebrate and invertebrate genomes. Each slice lists the number and percentages (in parentheses) of "strict orthologs" between the human, fly, and worm genomes involved in a given category of molecular function. "Strict orthologs" are defined here as bi-directional BLAST best hits (*180*) such that each orthologous pair (i) has a BLASTP *P*-value of $\leq 10^{-10}$ (*120*), and (ii) has a more significant BLASTP score than any paralogs in either organism, i.e., there has likely been no duplication subsequent to speciation that might make the orthology ambiguous. This measure is quite strict and is a lower bound on the number of orthologs. By these criteria, there are 2758 strict human-fly orthologs, and 2031 human-worm orthologs (1523 in common between these sets).

zymes are involved in intermediary metabolism. The only exception is the hydrolase category, which is not significantly overrepresented in the shared protein set. Proteases form the largest part of this category, and several large protease families have expanded in each of these three organisms after their divergence. The category of select regulatory molecules is also overrepresented in the conserved set. The major conserved families are small guanosine triphosphatases (GTPases) (especially the Ras-related superfamily, including ADP ribosylation factor) and cell cycle regulators (particularly the cullin family, cyclin C family, and several cell division protein kinases). The last two significantly overrepresented categories are protein transport and trafficking, and chaperones. The most conserved groups in these categories are proteins involved in coated vesicle-mediated transport, and chaperones involved in protein folding and heat-shock response [particularly the DNAJ family, and heat-shock protein 60 (HSP60), HSP70, and HSP90 families]. These observations provide only a conservative estimate of the protein families in the context of specific cellular processes that were likely derived from the last common ancestor of the human, fly, and worm. As stated before, this analysis does not provide a complete estimate of conservation across the three animal genomes, as paralogous duplication makes the determination of true orthologs difficult within the members of conserved protein families.

7.3 Differences between the human genome and other sequenced eukaryotic genomes

To explore the molecular building blocks of the vertebrate taxon, we have compared the human genome with the other sequenced eukaryotic genomes at three levels: molecular functions, protein families, and protein domains.

Molecular differences can be correlated with phenotypic differences to begin to reveal the developmental and cellular processes that are unique to the vertebrates. Tables 18 and 19 display a comparison among all sequenced eukaryotic genomes, over selected protein/domain families (defined by sequence similarity, e.g., the serine-threonine protein kinases) and superfamilies (defined by shared molecular function, which may include several sequence-related families, e.g., the cytokines). In these tables we have focused on (super) families that are either very large or that differ significantly in humans compared with the other sequenced eukaryote genomes. We have found that the most prominent human expansions are in proteins involved in (i) acquired immune functions; (ii) neural development, structure, and functions; (iii) intercellular and intracellular signaling pathways in development and homeostasis; (iv) hemostasis; and (v) apoptosis.

Acquired immunity. One of the most striking differences between the human genome and the *Drosophila* or *C. elegans* genome is the appearance of genes involved in acquired immunity (Tables 18 and 19). This is expected, because the acquired immune response is a defense system that only occurs in vertebrates. We observe 22 class I and 22 class II major histocompatibility complex (MHC) antigen genes and 114 other immunoglobulin genes in the human genome. In addition, there are 59 genes in the cognate immunoglobulin receptor family. At the domain level, this is exemplified by an expansion and recruitment of the ancient immunoglobulin fold to constitute molecules such as MHC, and of the integrin fold to form several of the cell adhesion molecules that mediate interactions between immune effector cells and the extracellular matrix. Vertebrate-specific proteins include the paracrine immune regulators family of secreted 4-alpha helical bundle proteins, namely the cytokines and chemokines. Some of the cytoplasmic signal transduction components associated with cytokine receptor signal transduction are also features that are poorly represented in the fly and worm. These include protein domains found in the signal transducer and activator of transcription (STATs), the suppressors of cytokine signaling (SOCS), and protein inhibitors of activated STATs (PIAS). In contrast, many of the animal-specific protein domains that play a role in innate immune response, such as the Toll receptors, do not appear to be significantly expanded in the human genome.

Neural development, structure, and function. In the human genome, as compared with the worm and fly genomes, there is a marked increase in the number of members of protein families that are involved in neural development. Examples include neurotrophic factors such as ependymin, nerve growth factor, and signaling molecules such as semaphorins, as well as the number of proteins involved directly in neural structure and function such as myelin proteins, voltage-gated ion channels, and synaptic proteins such as synaptotagmin. These observations correlate well with the known phenotypic differences between the nervous systems of these taxa, notably (i) the increase in the number and connectivity of neurons; (ii) the increase in number of distinct neural cell types (as many as a thousand or more in human compared with a few hundred in fly and worm) (*121*); (iii) the increased length of individual axons; and (iv) the significant increase in glial cell number, especially the appearance of myelinating glial cells, which are electrically inert supporting cells differentiated from the same stem cells as neurons. A number of prominent protein expansions are involved in the processes of neural development. Of the extracellular domains that mediate cell adhesion, the connexin domain–containing proteins (*122*) exist only in humans. These proteins, which are not present in the *Drosophila* or *C. elegans* genomes, appear to provide the constitutive subunits of intercellular channels and the structural basis for electrical coupling. Pathway finding by axons and neuronal network formation is mediated through a subset of ephrins and their cognate receptor tyrosine kinases that act as positional labels to establish topographical projections (*123*). The probable biological role for the semaphorins (22 in human compared with 6 in the fly and 2 in the worm) and their receptors (neuropilins and plexins) is that of axonal guidance molecules (*124*). Signaling molecules such as neurotrophic factors and some cytokines have been shown to regulate neuronal cell survival, proliferation, and axon guidance (*125*). Notch receptors and ligands play important roles in glial cell fate determination and gliogenesis (*126*).

Other human expanded gene families play key roles directly in neural structure and function. One example is synaptotagmin (expanded more than twofold in humans relative to the invertebrates), originally found to regulate synaptic transmission by serving as a Ca^{2+} sensor (or receptor) during synaptic vesicle fusion and release (*127*). Of interest is the increased co-occurrence in humans of PDZ and the SH3 domains in neuronal-specific adaptor molecules; examples include proteins that likely modulate channel activity at synaptic junctions (*128*). We also noted expansions in several ion-channel families (Table 19), including the EAG subfamily (related to cyclic nucleotide gated channels), the voltage-gated calcium/sodium channel family, the inward-rectifier potassium channel family, and the voltage-gated potassium channel, alpha subunit family. Voltage-gated sodium and potassium channels are involved in the generation of action potentials in neurons. Together with voltage-gated calcium channels, they also play a key role in coupling action potentials to neurotransmitter release, in the development of neurites, and in short-term memory. The recent observation of a calcium-regulated association between sodium channels and synaptotagmin may have consequences for the establishment and regulation of neuronal excitability (*129*).

Myelin basic protein and myelin-associated glycoprotein are major classes of protein components in both the central and peripheral nervous system of vertebrates. Myelin P0 is a major component of peripheral myelin, and myelin proteolipid and myelin oligodendrocyte glycopotein are found in the central nervous system. Mutations in any of these

Table 18. Domain-based comparative analysis of proteins in *H. sapiens* (H), *D. melanogaster* (F), *C. elegans* (W), *S. cerevisiae* (Y), and *A. thaliana* (A). The predicted protein set of each of the above eukaryotic organisms was analyzed with Pfam version 5.5 using E value cutoffs of 0.001. The number of proteins containing the specified Pfam domains as well as the total number of domains (in parentheses) are shown in each column. Domains were categorized into cellular processes for presentation. Some domains (i.e., SH2) are listed in more than one cellular process. Results of the Pfam analysis may differ from results obtained based on human curation of protein families, owing to the limitations of large-scale automatic classifications. Representative examples of domains with reduced counts owing to the stringent E value cutoff used for this analysis are marked with a double asterisk (**). Examples include short divergent and predominantly alpha-helical domains, and certain classes of cysteine-rich zinc finger proteins.

Accession number	Domain name	Domain description	H	F	W	Y	A
		Developmental and homeostatic regulators					
PF02039	Adrenomedullin	Adrenomedullin	1	0	0	0	0
PF00212	ANP	Atrial natriuretic peptide	2	0	0	0	0
PF00028	Cadherin	Cadherin domain	100 (550)	14 (157)	16 (66)	0	0
PF00214	Calc_CGRP_IAPP	Calcitonin/CGRP/IAPP family	3	0	0	0	0
PF01110	CNTF	Ciliary neurotrophic factor	1	0	0	0	0
PF01093	Clusterin	Clusterin	3	0	0	0	0
PF00029	Connexin	Connexin	14 (16)	0	0	0	0
PF00976	ACTH_domain	Corticotropin ACTH domain	1	0	0	0	0
PF00473	CRF	Corticotropin-releasing factor family	2	1	0	0	0
PF00007	Cys_knot	Cystine-knot domain	10 (11)	2	0	0	0
PF00778	DIX	Dix domain	5	2	4	0	0
PF00322	Endothelin	Endothelin family	3	0	0	0	0
PF00812	Ephrin	Ephrin	7 (8)	2	4	0	0
PF01404	EPh_lbd	Ephrin receptor ligand binding domain	12	2	1	0	0
PF00167	FGF	Fibroblast growth factor	23	1	1	0	0
PF01534	Frizzled	Frizzled/Smoothened family membrane region	9	7	3	0	0
PF00236	Hormone6	Glycoprotein hormones	1	0	0	0	0
PF01153	Glypican	Glypican	14	2	1	0	0
PF01271	Granin	Grainin (chromogranin or secretogranin)	3	0	0	0	0
PF02058	Guanylin	Guanylin precursor	1	0	0	0	0
PF00049	Insulin	Insulin/IGF/Relaxin family	7	4	0	0	0
PF00219	IGFBP	Insulin-like growth factor binding proteins	10	0	0	0	0
PF02024	Leptin	Leptin	1	0	0	0	0
PF00193	Xlink	LINK (hyaluron binding)	13 (23)	0	1	0	0
PF00243	NGF	Nerve growth factor family	3	0	0	0	0
PF02158	Neuregulin	Neuregulin family	4	0	0	0	0
PF00184	Hormone5	Neurohypophysial hormones	1	0	0	0	0
PF02070	NMU	Neuromedin U	1	0	0	0	0
PF00066	Notch	Notch (DSL) domain	3 (5)	2 (4)	2 (6)	0	0
PF00865	Osteopontin	Osteopontin	1	0	0	0	0
PF00159	Hormone3	Pancreatic hormone peptides	3	0	0	0	0
PF01279	Parathyroid	Parathyroid hormone family	2	0	0	0	0
PF00123	Hormone2	Peptide hormone	5 (9)	0	0	0	0
PF00341	PDGF	Platelet-derived growth factor (PDGF)	5	1	0	0	0
PF01403	Sema	Sema domain	27 (29)	8 (10)	3 (4)	0	0
PF01033	Somatomedin_B	Somatomedin B domain	5 (8)	3	0	0	0
PF00103	Hormone	Somatotropin	1	0	0	0	0
PF02208	Sorb	Sorbin homologous domain	2	0	0	0	0
PF02404	SCF	Stem cell factor	2	0	0	0	0
PF01034	Syndecan	Syndecan domain	3	1	1	0	0
PF00020	TNFR_c6	TNFR/NGFR cysteine-rich region	17 (31)	1	0	0	0
PF00019	TGF-β	Transforming growth factor β-like domain	27 (28)	6	4	0	0
PF01099	Uteroglobin	Uteroglobin family	3	0	0	0	0
PF01160	Opiods_neuropep	Vertebrate endogenous opioids neuropeptide	3	0	0	0	0
PF00110	Wnt	Wnt family of developmental signaling proteins	18	7 (10)	5	0	0
		Hemostasis					
PF01821	ANATO	Anaphylotoxin-like domain	6 (14)	0	0	0	0
PF00386	C1q	C1q domain	24	0	0	0	0
PF00200	Disintegrin	Disintegrin	18	2	3	0	0
PF00754	F5_F8_type_C	F5/8 type C domain	15 (20)	5 (6)	2	0	0
PF01410	COLFI	Fibrillar collagen C-terminal domain	10	0	0	0	0
PF00039	Fn1	Fibronectin type I domain	5 (18)	0	0	0	0
PF00040	Fn2	Fibronectin type II domain	11 (16)	0	0	0	0
PF00051	Kringle	Kringle domain	15 (24)	2	2	0	0
PF01823	MACPF	MAC/Perforin domain	6	0	0	0	0
PF00354	Pentaxin	Pentaxin family	9	0	0	0	0
PF00277	SAA_proteins	Serum amyloid A protein	4	0	0	0	0
PF00084	Sushi	Sushi domain (SCR repeat)	53 (191)	11 (42)	8 (45)	0	0
PF02210	TSPN	Thrombospondin N-terminal–like domains	14	1	0	0	0
PF01108	Tissue_fac	Tissue factor	1	0	0	0	0
PF00868	Transglutamin_N	Transglutaminase family	6	1	0	0	0
PF00927	Transglutamin_C	Transglutaminase family	8	1	0	0	0

Table 18 (*Continued*)

Accession number	Domain name	Domain description	H	F	W	Y	A
PF00594	Gla	Vitamin K-dependent carboxylation/gamma-carboxyglutamic (GLA) domain	11	0	0	0	0
		Immune response					
PF00711	Defensin_beta	Beta defensin	1	0	0	0	0
PF00748	Calpain_inhib	Calpain inhibitor repeat	3 (9)	0	0	0	0
PF00666	Cathelicidins	Cathelicidins	2	0	0	0	0
PF00129	MHC_I	Class I histocompatibility antigen, domains alpha 1 and 2	18 (20)	0	0	0	0
PF00993	MHC_II_alpha**	Class II histocompatibility antigen, alpha domain	5 (6)	0	0	0	0
PF00969	MHC_II_beta**	Class II histocompatibility antigen, beta domain	7	0	0	0	0
PF00879	Defensin_propep	Defensin propeptide	3	0	0	0	0
PF01109	GM_CSF	Granulocyte-macrophage colony-stimulating factor	1	0	0	0	0
PF00047	Ig	Immunoglobulin domain	381 (930)	125 (291)	67 (323)	0	0
PF00143	Interferon	Interferon alpha/beta domain	7 (9)	0	0	0	0
PF00714	IFN-gamma	Interferon gamma	1	0	0	0	0
PF00726	IL10	Interleukin-10	1	0	0	0	0
PF02372	IL15	Interleukin-15	1	0	0	0	0
PF00715	IL2	Interleukin-2	1	0	0	0	0
PF00727	IL4	Interleukin-4	1	0	0	0	0
PF02025	IL5	Interleukin-5	1	0	0	0	0
PF01415	IL7	Interleukin-7/9 family	1	0	0	0	0
PF00340	IL1	Interleukin-1	7	0	0	0	0
PF02394	IL1_propep	Interleukin-1 propeptide	1	0	0	0	0
PF02059	IL3	Interleukin-3	1	0	0	0	0
PF00489	IL6	Interleukin-6/G-CSF/MGF family	2	0	0	0	0
PF01291	LIF_OSM	Leukemia inhibitory factor (LIF)/oncostatin (OSM) family	2	0	0	0	0
PF00323	Defensins	Mammalian defensin	2	0	0	0	0
PF01091	PTN_MK	PTN/MK heparin-binding protein	2	0	0	0	0
PF00277	SAA_proteins	Serum amyloid A protein	4	0	0	0	0
PF00048	IL8	Small cytokines (intecrine/chemokine), interleukin-8 like	32	0	0	0	0
[illegible]01582	TIR	TIR domain	18	8	2	0	131 (143)
[illegible]00229	TNF	TNF (tumor necrosis factor) family	12	0	0	0	0
PF00088	Trefoil	Trefoil (P-type) domain	5 (6)	0	2	0	0
		PI-PY-rho GTPase signaling					
PF00779	BTK	BTK motif	5	1	0	0	0
PF00168	C2	C2 domain	73 (101)	32 (44)	24 (35)	6 (9)	66 (90)
PF00609	DAGKa	Diacylglycerol kinase accessory domain (presumed)	9	4	7	0	6
PF00781	DAGKc	Diacylglycerol kinase catalytic domain (presumed)	10	8	8	2	11 (12)
PF00610	DEP	Domain found in Dishevelled, Egl-10, and Pleckstrin (DEP)	12 (13)	4	10	5	2
PF01363	FYVE	FYVE zinc finger	28 (30)	14	15	5	15
PF00996	GDI	GDP dissociation inhibitor	6	2	1	1	3
PF00503	G-alpha	G-protein alpha subunit	27 (30)	10	20 (23)	2	5
PF00631	G-gamma	G-protein gamma like domains	16	5	5	1	0
PF00616	RasGAP	GTPase-activator protein for Ras-like GTPase	11	5	8	3	0
PF00618	RasGEFN	Guanine nucleotide exchange factor for Ras-like GTPases; N-terminal motif	9	2	3	5	0
PF00625	Guanylate_kin	Guanylate kinase	12	8	7	1	4
PF02189	ITAM	Immunoreceptor tyrosine-based activation motif	3	0	0	0	0
PF00169	PH	PH domain	193 (212)	72 (78)	65 (68)	24	23
PF00130	DAG_PE-bind	Phorbol esters/diacylglycerol binding domain (C1 domain)	45 (56)	25 (31)	26 (40)	1 (2)	4
PF00388	PI-PLC-X	Phosphatidylinositol-specific phospholipase C, X domain	12	3	7	1	8
PF00387	PI-PLC-Y	Phosphatidylinositol-specific phospholipase C, Y domain	11	2	7	1	8
PF00640	PID	Phosphotyrosine interaction domain (PTB/PID)	24 (27)	13	11 (12)	0	0
PF02192	PI3K_p85B	PI3-kinase family, p85-binding domain	2	1	1	0	0
PF00794	PI3K_rbd	PI3-kinase family, ras-binding domain	6	3	1	0	0
PF01412	ArfGAP	Putative GTP-ase activating protein for Arf	16	9	8	6	15
PF02196	RBD	Raf-like Ras-binding domain	6 (7)	4	1	0	0
PF02145	Rap_GAP	Rap/ran-GAP	5	4	2	0	0
PF00788	RA	Ras association (RalGDS/AF-6) domain	18 (19)	7 (9)	6	1	0
[illegible]00071	Ras	Ras family	126	56 (57)	51	23	78
[illegible]0617	RasGEF	RasGEF domain	21	8	7	5	0
[illegible]00615	RGS	Regulator of G protein signaling domain	27	6 (7)	12 (13)	1	0
PF02197	RIIa	Regulatory subunit of type II PKA R-subunit	4	1	2	1	0

Table 18 (*Continued*)

Accession number	Domain name	Domain description	H	F	W	Y	A
PF00620	RhoGAP	RhoGAP domain	59	19	20	9	8
PF00621	RhoGEF	RhoGEF domain	46	23 (24)	18 (19)	3	0
PF00536	SAM	SAM domain (Sterile alpha motif)	29 (31)	15	8	3	6
PF01369	Sec7	Sec7 domain	13	5	5	5	9
PF00017	SH2	Src homology 2 (SH2) domain	87 (95)	33 (39)	44 (48)	1	3
PF00018	SH3	Src homology 3 (SH3) domain	143 (182)	55 (75)	46 (61)	23 (27)	4
PF01017	STAT	STAT protein	7	1	1 (2)	0	0
PF00790	VHS	VHS domain	4	2	4	4	8
PF00568	WH1	WH1 domain	7	2	2 (3)	1	0
		Domains involved in apoptosis					
PF00452	Bcl-2	Bcl-2	9	2	1	0	0
PF02180	BH4	Bcl-2 homology region 4	3	0	1	0	0
PF00619	CARD	Caspase recruitment domain	16	0	2	0	0
PF00531	Death	Death domain	16	5	7	0	0
PF01335	DED	Death effector domain	4 (5)	0	0	0	0
PF02179	BAG	Domain present in Hsp70 regulators	5 (8)	3	2	1	5
PF00656	ICE_p20	ICE-like protease (caspase) p20 domain	11	7	3	0	0
PF00653	BIR	Inhibitor of Apoptosis domain	8 (14)	5 (9)	2 (3)	1 (2)	0
		Cytoskeletal					
PF00022	Actin	Actin	61 (64)	15 (16)	12	9 (11)	24
PF00191	Annexin	Annexin	16 (55)	4 (16)	4 (11)	0	6 (16)
PF00402	Calponin	Calponin family	13 (22)	3	7 (19)	0	0
PF00373	Band_41	FERM domain (Band 4.1 family)	29 (30)	17 (19)	11 (14)	0	0
PF00880	Nebulin_repeat	Nebulin repeat	4 (148)	1 (2)	1	0	0
PF00681	Plectin_repeat	Plectin repeat	2 (11)	0	0	0	0
PF00435	Spectrin	Spectrin repeat	31 (195)	13 (171)	10 (93)	0	0
PF00418	Tubulin-binding	Tau and MAP proteins, tubulin-binding	4 (12)	1 (4)	2 (8)	0	0
PF00992	Troponin	Troponin	4	6	8	0	0
PF02209	VHP	Villin headpiece domain	5	2	2	0	5
PF01044	Vinculin	Vinculin family	4	2	1	0	0
		ECM adhesion					
PF01391	Collagen	Collagen triple helix repeat (20 copies)	65 (279)	10 (46)	174 (384)	0	0
PF01413	C4	C-terminal tandem repeated domain in type 4 procollagen	6 (11)	2 (4)	3 (6)	0	0
PF00431	CUB	CUB domain	47 (69)	9 (47)	43 (67)	0	0
PF00008	EGF	EGF-like domain	108 (420)	45 (186)	54 (157)	0	1
PF00147	Fibrinogen_C	Fibrinogen beta and gamma chains, C-terminal globular domain	26	10 (11)	6	0	0
PF00041	Fn3	Fibronectin type III domain	106 (545)	42 (168)	34 (156)	0	1
PF00757	Furin-like	Furin-like cysteine rich region	5	2	1	0	0
PF00357	Integrin_A	Integrin alpha cytoplasmic region	3	1	2	0	0
PF00362	Integrin_B	Integrins, beta chain	8	2	2	0	0
PF00052	Laminin_B	Laminin B (Domain IV)	8 (12)	4 (7)	6 (10)	0	0
PF00053	Laminin_EGF	Laminin EGF-like (Domains III and V)	24 (126)	9 (62)	11 (65)	0	0
PF00054	Laminin_G	Laminin G domain	30 (57)	18 (42)	14 (26)	0	0
PF00055	Laminin_Nterm	Laminin N-terminal (Domain VI)	10	6	4	0	0
PF00059	Lectin_c	Lectin C-type domain	47 (76)	23 (24)	91 (132)	0	0
PF01463	LRRCT	Leucine rich repeat C-terminal domain	69 (81)	23 (30)	7 (9)	0	0
PF01462	LRRNT	Leucine rich repeat N-terminal domain	40 (44)	7 (13)	3 (6)	0	0
PF00057	Ldl_recept_a	Low-density lipoprotein receptor domain class A	35 (127)	33 (152)	27 (113)	0	0
PF00058	Ldl_recept_b	Low-density lipoprotein receptor repeat class B	15 (96)	9 (56)	7 (22)	0	0
PF00530	SRCR	Scavenger receptor cysteine-rich domain	11 (46)	4 (8)	1 (2)	0	0
PF00084	Sushi	Sushi domain (SCR repeat)	53 (191)	11 (42)	8 (45)	0	0
PF00090	Tsp_1	Thrombospondin type 1 domain	41 (66)	11 (23)	18 (47)	0	0
PF00092	Vwa	von Willebrand factor type A domain	34 (58)	0	17 (19)	0	1
PF00093	Vwc	von Willebrand factor type C domain	19 (28)	6 (11)	2 (5)	0	0
PF00094	Vwd	von Willebrand factor type D domain	15 (35)	3 (7)	9	0	0
		Protein interaction domains					
PF00244	14-3-3	14-3-3 proteins	20	3	3	2	15
PF00023	Ank	Ank repeat	145 (404)	72 (269)	75 (223)	12 (20)	66 (111)
PF00514	Armadillo_seg	Armadillo/beta-catenin-like repeats	22 (56)	11 (38)	3 (11)	2 (10)	25 (67)
PF00168	C2	C2 domain	73 (101)	32 (44)	24 (35)	6 (9)	66 (90)
PF00027	cNMP_binding	Cyclic nucleotide-binding domain	26 (31)	21 (33)	15 (20)	2 (3)	22
PF01556	DnaJ_C	DnaJ C terminal region	12	9	5	3	19
PF00226	DnaJ	DnaJ domain	44	34	33	20	93
PF00036	Efhand**	EF hand	83 (151)	64 (117)	41 (86)	4 (11)	120 (328)
PF00611	FCH	Fes/CIP4 homology domain	9	3	2	4	0
PF01846	FF	FF domain	4 (11)	4 (10)	3 (16)	2 (5)	4 (8)
PF00498	FHA	FHA domain	13	15	7	13 (14)	17

myelin proteins result in severe demyelination, which is a pathological condition in which the myelin is lost and the nerve conduction is severely impaired (*130*). Humans have at least 10 genes belonging to four different families involved in myelin production (five myelin P0, three myelin proteolipid, myelin basic protein, and myelin-oligodendrocyte glycoprotein, or MOG), and possibly more-remotely related members of the MOG family. Flies have only a single myelin proteolipid, and worms have none at all.

Intercellular and intracellular signaling pathways in development and homeostasis. Many protein families that have expanded in humans relative to the invertebrates are involved in signaling processes, particularly in response to development and differentiation

Table 18 (*Continued*)

Accession number	Domain name	Domain description	H	F	W	Y	A
PF00254	FKBP	FKBP-type peptidyl-prolyl cis-trans isomerases	15 (20)	7 (8)	7 (13)	4	24 (29)
PF01590	GAF	GAF domain	7 (8)	2 (4)	1	0	10
PF01344	Kelch	Kelch motif	54 (157)	12 (48)	13 (41)	3	102 (178)
PF00560	LRR**	Leucine Rich Repeat	25 (30)	24 (30)	7 (11)	1	15 (16)
PF00917	MATH	MATH domain	11	5	88 (161)	1	61 (74)
PF00989	PAS	PAS domain	18 (19)	9 (10)	6	1	13 (18)
PF00595	PDZ	PDZ domain (Also known as DHR or GLGF)	96 (154)	60 (87)	46 (66)	2	5
PF00169	PH	PH domain	193 (212)	72 (78)	65 (68)	24	23
PF01535	PPR**	PPR repeat	5	3 (4)	0	1	474 (2485)
PF00536	SAM	SAM domain (Sterile alpha motif)	29 (31)	15	8	3	6
PF01369	Sec7	Sec7 domain	13	5	5	5	9
PF00017	SH2	Src homology 2 (SH2) domain	87 (95)	33 (39)	44 (48)	1	3
PF00018	SH3	Src homology 3 (SH3) domain	143 (182)	55 (75)	46 (61)	23 (27)	4
PF01740	STAS	STAS domain	5	1	6	2	13
PF00515	TPR**	TPR domain	72 (131)	39 (101)	28 (54)	16 (31)	65 (124)
PF00400	WD40**	WD40 domain	136 (305)	98 (226)	72 (153)	56 (121)	167 (344)
PF00397	WW	WW domain	32 (53)	24 (39)	16 (24)	5 (8)	11 (15)
PF00569	ZZ	ZZ-Zinc finger present in dystrophin, CBP/p300	10 (11)	13	10	2	10
		Nuclear interaction domains					
PF01754	Zf-A20	A20-like zinc finger	2 (8)	2	2	0	8
PF01388	ARID	ARID DNA binding domain	11	6	4	2	7
PF01426	BAH	BAH domain	8 (10)	7 (8)	4 (5)	5	21 (25)
PF00643	Zf-B_box**	B-box zinc finger	32 (35)	1	2	0	0
PF00533	BRCT	BRCA1 C Terminus (BRCT) domain	17 (28)	10 (18)	23 (35)	10 (16)	12 (16)
PF00439	Bromodomain	Bromodomain	37 (48)	16 (22)	18 (26)	10 (15)	28
PF00651	BTB	BTB/POZ domain	97 (98)	62 (64)	86 (91)	1 (2)	30 (31)
PF00145	DNA_methylase	C-5 cytosine-specific DNA methylase	3 (4)	1	0	0	13 (15)
PF00385	Chromo	chromo' (CHRromatin Organization MOdifier) domain	24 (27)	14 (15)	17 (18)	1 (2)	12
PF00125	Histone	Core histone H2A/H2B/H3/H4	75 (81)	5	71 (73)	8	48
PF00134	Cyclin	Cyclin	19	10	10	11	35
PF00270	DEAD	DEAD/DEAH box helicase	63 (66)	48 (50)	55 (57)	50 (52)	84 (87)
PF01529	Zf-DHHC	DHHC zinc finger domain	15	20	16	7	22
PF00646	F-box**	F-box domain	16	15	309 (324)	9	165 (167)
PF00250	Fork_head	Fork head domain	35 (36)	20 (21)	15	4	0
PF00320	GATA	GATA zinc finger	11 (17)	5(6)	8 (10)	9	26
PF01585	G-patch	G-patch domain	18	16	13	4	14 (15)
PF00010	HLH**	Helix-loop-helix DNA-binding domain	60 (61)	44	24	4	39
PF00850	Hist_deacetyl	Histone deacetylase family	12	5 (6)	8 (10)	5	10
PF00046	Homeobox	Homeobox domain	160 (178)	100 (103)	82 (84)	6	66
PF01833	TIG	IPT/TIG domain	29 (53)	11 (13)	5 (7)	2	1
PF02373	JmjC	JmjC domain	10	4	6	4	7
PF02375	JmjN	JmjN domain	7	4	2	3	7
PF00013	KH-domain	KH domain	28 (67)	14 (32)	17 (46)	4 (14)	27 (61)
PF01352	KRAB	KRAB box	204 (243)	0	0	0	0
PF00104	Hormone_rec	Ligand-binding domain of nuclear hormone receptor	47	17	142 (147)	0	0
PF00412	LIM	LIM domain containing proteins	62 (129)	33 (83)	33 (79)	4 (7)	10 (16)
PF00917	MATH	MATH domain	11	5	88 (161)	1	61 (74)
PF00249	Myb_DNA-binding	Myb-like DNA-binding domain	32 (43)	18 (24)	17 (24)	15 (20)	243 (401)
PF02344	Myc-LZ	Myc leucine zipper domain	1	0	0	0	0
PF01753	Zf-MYND	MYND finger	14	14	9	1	7
PF00628	PHD	PHD-finger	68 (86)	40 (53)	32 (44)	14 (15)	96 (105)
PF00157	Pou	Pou domain—N-terminal to homeobox domain	15	5	4	0	0
PF02257	RFX_DNA_binding	RFX DNA-binding domain	7	2	1	1	0
PF00076	Rrm	RNA recognition motif (a.k.a. RRM, RBD, or RNP domain)	224 (324)	127 (199)	94 (145)	43 (73)	232 (369)
PF02037	SAP	SAP domain	15	8	5	5	6 (7)
PF00622	SPRY	SPRY domain	44 (51)	10 (12)	5 (7)	3	6
PF01852	START	START domain	10	2	6	0	23
PF00907	T-box	T-box	17 (19)	8	22	0	0

Table 18 (*Continued*)

Accession number	Domain name	Domain description	H	F	W	Y	A
PF02135	Zf-TAZ	TAZ finger	2 (3)	1 (2)	6 (7)	0	10 (15)
PF01285	TEA	TEA domain	4	1	1	1	0
PF02176	Zf-TRAF	TRAF-type zinc finger	6 (9)	1 (3)	1	0	2
PF00352	TBP	Transcription factor TFIID (or TATA-binding protein, TBP)	2 (4)	4 (8)	2 (4)	1 (2)	2 (4)
PF00567	TUDOR	TUDOR domain	9 (24)	9 (19)	4 (5)	0	2
PF00642	Zf-CCCH	Zinc finger C-x8-C-x5-C-x3-H type (and similar)	17 (22)	6 (8)	22 (42)	3 (5)	31 (46)
PF00096	Zf-C2H2**	Zinc finger, C2H2 type	564 (4500)	234 (771)	68 (155)	34 (56)	21 (24)
PF00097	Zf-C3HC4	Zinc finger, C3HC4 type (RING finger)	135 (137)	57	88 (89)	18	298 (304)
PF00098	Zf-CCHC	Zinc knuckle	9 (17)	6 (10)	17 (33)	7 (13)	68 (91)

(Tables 18 and 19). They include secreted hormones and growth factors, receptors, intracellular signaling molecules, and transcription factors.

Developmental signaling molecules that are enriched in the human genome include growth factors such as wnt, transforming growth factor–β (TGF-β), fibroblast growth factor (FGF), nerve growth factor, platelet derived growth factor (PDGF), and ephrins. These growth factors affect tissue differentiation and a wide range of cellular processes involving actin-cytoskeletal and nuclear regulation. The corresponding receptors of these developmental ligands are also expanded in humans. For example, our analysis suggests at least 8 human ephrin genes (2 in the fly, 4 in the worm) and 12 ephrin receptors (2 in the fly, 1 in the worm). In the wnt signaling pathway, we find 18 wnt family genes (6 in the fly, 5 in the worm) and 12 frizzled receptors (6 in the fly, 5 in the worm). The Groucho family of transcriptional corepressors downstream in the wnt pathway are even more markedly expanded, with 13 predicted members in humans (2 in the fly, 1 in the worm).

Extracellular adhesion molecules involved in signaling are expanded in the human genome (Tables 18 and 19). The interactions of several of these adhesion domains with extracellular matrix proteoglycans play a critical role in host defense, morphogenesis, and tissue repair (*131*). Consistent with the well-defined role of heparan sulfate proteoglycans in modulating these interactions (*132*), we observe an expansion of the heparin sulfate sulfotransferases in the human genome relative to worm and fly. These sulfotransferases modulate tissue differentiation (*133*). A similar expansion in humans is noted in structural proteins that constitute the actin-cytoskeletal architecture. Compared with the fly and worm, we observe an explosive expansion of the nebulin (35 domains per protein on average), aggrecan (12 domains per protein on average), and plectin (5 domains per protein on average) repeats in humans. These repeats are present in proteins involved in modulating the actin-cytoskeleton with predominant expression in neuronal, muscle, and vascular tissues.

Comparison across the five sequenced eukaryotic organisms revealed several expanded protein families and domains involved in cytoplasmic signal transduction (Table 18). In particular, signal transduction pathways playing roles in developmental regulation and acquired immunity were substantially enriched. There is a factor of 2 or greater expansion in humans in the Ras superfamily GTPases and the GTPase activator and GTP exchange factors associated with them. Although there are about the same number of tyrosine kinases in the human and *C. elegans* genomes, in humans there is an increase in the SH2, PTB, and ITAM domains involved in phosphotyrosine signal transduction. Further, there is a twofold expansion of phosphodiesterases in the human genome compared with either the worm or fly genomes.

The downstream effectors of the intracellular signaling molecules include the transcription factors that transduce developmental fates. Significant expansions are noted in the ligand-binding nuclear hormone receptor class of transcription factors compared with the fly genome, although not to the extent observed in the worm (Tables 18 and 19). Perhaps the most striking expansion in humans is in the C2H2 zinc finger transcription factors. Pfam detects a total of 4500 C2H2 zinc finger domains in 564 human proteins, compared with 771 in 234 fly proteins. This means that there has been a dramatic expansion not only in the number of C2H2 transcription factors, but also in the number of these DNA-binding motifs per transcription factor (8 on average in humans, 3.3 on average in the fly, and 2.3 on average in the worm). Furthermore, many of these transcription factors contain either the KRAB or SCAN domains, which are not found in the fly or worm genomes. These domains are involved in the oligomerization of transcription factors and increase the combinatorial partnering of these factors. In general, most of the transcription factor domains are shared between the three animal genomes, but the reassortment of these domains results in organism-specific transcription factor families. The domain combinations found in the human, fly, and worm include the BTB with C2H2 in the fly and humans, and homeodomains alone or in combination with Pou and LIM domains in all of the animal genomes. In plants, however, a different set of transcription factors are expanded, namely, the myb family, and a unique set that includes VP1 and AP2 domain–containing proteins (*134*). The yeast genome has a paucity of transcription factors compared with the multicellular eukaryotes, and its repertoire is limited to the expansion of the yeast-specific C6 transcription factor family involved in metabolic regulation.

While we have illustrated expansions in a subset of signal transduction molecules in the human genome compared with the other eukaryotic genomes, it should be noted that most of the protein domains are highly conserved. An interesting observation is that worms and humans have approximately the same number of both tyrosine kinases and serine/threonine kinases (Table 19). It is important to note, however, that these are merely counts of the catalytic domain; the proteins that contain these domains also display a wide repertoire of interaction domains with significant combinatorial diversity.

Hemostasis. Hemostasis is regulated primarily by plasma proteases of the coagulation pathway and by the interactions that occur between the vascular endothelium and platelets. Consistent with known anatomical and physiological differences between vertebrates and invertebrates, extracellular adhesion domains that constitute proteins integral to hemostasis are expanded in the human relative to the fly and worm (Tables 18 and 19). We note the evolution of domains such as FIMAC, FN1, FN2, and C1q that mediate surface interactions between hematopoeitic cells and the vascular matrix. In addition, there has been extensive recruitment of more-ancient animal-specific domains such as VWA, VWC, VWD, kringle, and FN3 into multidomain proteins that are involved in hemostatic regulation. Although we do not find a large expansion in the total number of serine proteases, this enzymatic domain has been specifically recruited into several of these multidomain proteins for proteolytic regulation in the vascular compartment. These are represented in plasma proteins that belong to the kinin and complement pathways. There is a

significant expansion in two families of matrix metalloproteases: ADAM (a disintegrin and metalloprotease) and MMPs (matrix metalloproteases) (Table 19). Proteolysis of extracellular matrix (ECM) proteins is critical for tissue development and for tissue degradation in diseases such as cancer, arthritis, Alzheimer's disease, and a variety of inflammatory conditions (*135*, *136*). ADAMs are a family of integral membrane proteins with a pivotal role in fibrinogenolysis and modulating interactions between hematopoietic components and the vascular matrix components. These proteins have been shown to cleave matrix proteins, and even signaling molecules: ADAM-17 converts tumor necrosis factor–α, and ADAM-10 has been implicated in the Notch signaling pathway (*135*). We have identified 19 members of the matrix metalloprotease family, and a total of 51 members of the ADAM and ADAM-TS families.

Apoptosis. Evolutionary conservation of some of the apoptotic pathway components across eukarya is consistent with its central role in developmental regulation and as a response to pathogens and stress signals. The signal transduction pathways involved in programmed cell death, or apoptosis, are mediated by interactions between well-characterized domains that include extracellular domains, adaptor (protein-protein interaction) domains, and those found in effector and regulatory enzymes (*137*). We enumerated the protein counts of central adaptor and effector enzyme domains that are found only in the apoptotic pathways to provide an estimate of divergence across eukarya and relative expansion in the human genome when compared with the fly and worm (Table 18). Adaptor domains found in proteins restricted only to apoptotic regulation such as the DED domains are vertebrate-specific, whereas others like BIR, CARD, and Bcl2 are represented in the fly and worm (although the number of Bcl2 family members in humans is significantly expanded). Although plants and yeast lack the caspases, caspase-like molecules, namely the para- and meta-caspases, have been reported in these organisms (*138*). Compared with other animal genomes, the human genome shows an expansion in the adaptor and effector domain–containing proteins involved in apoptosis, as well as in the proteases involved in the cascade such as the caspase and calpain families.

Expansions of other protein families. *Metabolic enzymes.* There are fewer cytochrome P450 genes in humans than in either the fly or worm. Lipoxygenases (six in humans), on the other hand, appear to be specific to the vertebrates and plants, whereas the lipoxygenase-activating proteins (four in humans) may be vertebrate-specific. Lipoxygenases are involved in arachidonic acid metabolism, and they and their activators have been implicated in diverse human pathology ranging from allergic responses to cancers. One of the most surprising human expansions, however, is in the number of glyceraldehyde-3-phosphate dehydrogenase (GAPDH) genes (46 in humans, 3 in the fly, and 4 in the worm). There is, however, evidence for many retrotransposed GAPDH pseudogenes (*139*), which may account for this apparent expansion. However, it is interesting that GAPDH, long known as a conserved enzyme involved in basic metabolism found across all phyla from bacteria to humans, has recently been shown to have other functions. It has a second cat-

Table 19. Number of proteins assigned to selected Panther families or subfamilies in *H. sapiens* (H), *D. melanogaster* (F), *C. elegans* (W), *S. cerevisiae* (Y), and *A. thaliana* (A).

Panther family/subfamily*	H	F	W	Y	A
Neural structure, function, development					
Ependymin	1	0	0	0	0
Ion channels					
Acetylcholine receptor	17	12	56	0	0
Amiloride-sensitive/degenerin	11	24	27	0	0
CNG/EAG	22	9	9	0	30
IRK	16	3	3	0	0
ITP/ryanodine	10	2	4	0	0
Neurotransmitter-gated	61	51	59	0	19
P2X purinoceptor	10	0	0	0	0
TASK	12	12	48	1	5
Transient receptor	15	3	3	1	0
Voltage-gated Ca^{2+} alpha	22	4	8	2	2
Voltage-gated Ca^{2+} alpha-2	10	3	2	0	0
Voltage-gated Ca^{2+} beta	5	2	2	0	0
Voltage-gated Ca^{2+} gamma	1	0	0	0	0
Voltage-gated K^+ alpha	33	5	11	0	0
Voltage-gated KQT	6	2	3	0	0
Voltage-gated Na^+	11	4	4	9	1
Myelin basic protein	1	0	0	0	0
Myelin PO	5	0	0	0	0
Myelin proteolipid	3	1	0	0	0
Myelin-oligodendrocyte glycoprotein	1	0	0	0	0
Neuropilin	2	0	0	0	0
Plexin	9	2	0	0	0
Semaphorin	22	6	2	0	0
Synaptotagmin	10	3	3	0	0
Immune response					
Defensin	3	0	0	0	0
Cytokine†	86	14	1	0	0
GCSF	1	0	0	0	0
GMCSF	1	0	0	0	0
Intercrine alpha	15	0	0	0	0
Intercrine beta	5	0	0	0	0
Inteferon	8	0	0	0	0
Interleukin	26	1	1	0	0
Leukemia inhibitory factor	1	0	0	0	0
MCSF	1	0	0	0	0
Peptidoglycan recognition protein	2	13	0	0	0
Pre-B cell enhancing factor	1	0	0	0	0
Small inducible cytokine A	14	0	0	0	0
Sl cytokine	2	0	0	0	0
TNF	9	0	0	0	0
Cytokine receptor†	62	1	0	0	0
Bradykinin/C-C chemokine receptor	7	0	0	0	0
Fl cytokine receptor	2	0	0	0	0
Interferon receptor	3	0	0	0	0
Interleukin receptor	32	0	0	0	0
Leukocyte tyrosine kinase receptor	3	0	0	0	0
MCSF receptor	1	0	0	0	0
TNF receptor	3	0	0	0	0
Immunoglobulin receptor†	59	0	0	0	0
T-cell receptor alpha chain	16	0	0	0	0
T-cell receptor beta chain	15	0	0	0	0
T-cell receptor gamma chain	1	0	0	0	0
T-cell receptor delta chain	1	0	0	0	0
Immunoglobulin FC receptor	8	0	0	0	0
Killer cell receptor	16	0	0	0	0
Polymeric-immunoglobulin receptor	4	0	0	0	0

alytic activity, as a uracil DNA glycosylase (*140*) and functions as a cell cycle regulator (*141*) and has even been implicated in apoptosis (*142*).

Translation. Another striking set of human expansions has occurred in certain families involved in the translational machinery. We identified 28 different ribosomal subunits that each have at least 10 copies in the genome; on average, for all ribosomal proteins there is about an 8- to 10-fold expansion in the number of genes relative to either the worm or fly. Retrotransposed pseudogenes may account for many of these expansions [see the discussion above and (*143*)]. Recent evidence suggests that a number of ribosomal proteins have secondary functions independent of their involvement in protein biosynthesis; for example, L13a and the related L7 subunits (36 copies in humans) have been shown to induce apoptosis (*144*).

There is also a four- to fivefold expansion in the elongation factor 1-alpha family (eEF1A; 56 human genes). Many of these expansions likely represent intronless paralogs that have presumably arisen from retrotransposition, and again there is evidence that many of these may be pseudogenes (*145*). However, a second form (eEF1A2) of this factor has been identied with tissue-specific expression in skeletal muscle and a complementary expression pattern to the ubiquitously expressed eEF1A (*146*).

Ribonucleoproteins. Alternative splicing results in multiple transcripts from a single gene, and can therefore generate additional diversity in an organism's protein complement. We have identified 269 genes for ribonucleoproteins. This represents over 2.5 times the number of ribonucleoprotein genes in the worm, two times that of the fly, and about the same as the 265 identified in the *Arabidopsis* genome. Whether the diversity of ribonucleoprotein genes in humans contributes to gene regulation at either the splicing or translational level is unknown.

Posttranslational modifications. In this set of processes, the most prominent expansion is the transglutaminases, calcium-dependent enzymes that catalyze the cross-linking of proteins in cellular processes such as hemostasis and apoptosis (*147*). The vitamin K–dependent gamma carboxylase gene product acts on the GLA domain (missing in the fly and worm) found in coagulation factors, osteocalcin, and matrix GLA protein (*148*). Tyrosylprotein sulfotransferases participate in the posttranslational modification of proteins involved in inflammation and hemostasis, including coagulation factors and chemokine receptors (*149*). Although there is no significant numerical increase in the counts for domains involved in nuclear protein modification, there are a number of domain arrangements in the predicted human proteins that are not found in the other currently sequenced genomes. These include the tandem association of two histone deacetylase domains in HD6 with a ubiquitin finger domain, a feature lacking in the fly genome. An additional example is the co-occurrence of important nuclear regulatory enzyme PARP (poly-ADP ribosyl transferase) domain fused to protein-interaction domains—BRCT and VWA in humans.

Concluding remarks. There are several possible explanations for the differences in phenotypic complexity observed in humans when compared to the fly and worm. Some of these relate to the prominent differences in the immune system, hemostasis, neuronal, vascular, and cytoskeletal complexity. The finding that the human genome contains fewer genes than previously predicted might be compensated for by combinatorial diversity generated at the levels of protein architecture, transcriptional and translational control, posttranslational modification of proteins, or posttranscriptional regulation. Extensive domain shuffling to increase or alter combinatorial diversity can provide an exponential

Table 19 (*Continued*)

Panther family/subfamily*	H	F	W	Y	A
MHC class I	22	0	0	0	0
MHC class II	20	0	0	0	0
Other immunoglobulin†	114	0	0	0	0
Toll receptor–related	10	6	0	0	0
Developmental and homeostatic regulators					
Signaling molecules†					
Calcitonin	3	0	0	0	0
Ephrin	8	2	4	0	0
FGF	24	1	1	0	0
Glucagon	4	0	0	0	0
Glycoprotein hormone beta chain	2	0	0	0	0
Insulin	1	0	0	0	0
Insulin-like hormone	3	0	0	0	0
Nerve growth factor	3	0	0	0	0
Neuregulin/heregulin	6	0	0	0	0
neuropeptide Y	4	0	0	0	0
PDGF	1	1	0	0	0
Relaxin	3	0	0	0	0
Stannocalcin	2	0	0	0	0
Thymopoeitin	2	0	1	0	0
Thyomosin beta	4	2	0	0	0
TGF-β	29	6	4	0	0
VEGF	4	0	0	0	0
Wnt	18	6	5	0	0
Receptors†					
Ephrin receptor	12	2	1	0	0
FGF receptor	4	4	0	0	0
Frizzled receptor	12	6	5	0	0
Parathyroid hormone receptor	2	0	0	0	0
VEGF receptor	5	0	0	0	0
BDNF/NT-3 nerve growth factor receptor	4	0	0	0	0
Kinases and phosphatases					
Dual-specificity protein phosphatase	29	8	10	4	11
S/T and dual-specificity protein kinase†	395	198	315	114	1102
S/T protein phosphatase	15	19	51	13	29
Y protein kinase†	106	47	100	5	16
Y protein phosphatase	56	22	95	5	6
Signal transduction					
ARF family	55	29	27	12	45
Cyclic nucleotide phosphodiesterase	25	8	6	1	0
G protein-coupled receptors†‡	616	146	284	0	1
G-protein alpha	27	10	22	2	5
G-protein beta	5	3	2	1	1
G-protein gamma	13	2	2	0	0
Ras superfamily	141	64	62	26	86
G-protein modulators†					
ARF GTPase-activating	20	8	9	5	15
Neurofibromin	7	2	0	2	0
Ras GTPase-activating	9	3	8	1	0
Tuberin	7	3	2	0	0
Vav proto-oncogene family	35	15	13	3	0

increase in the ability to mediate protein-protein interactions without dramatically increasing the absolute size of the protein complement (*150*). Evolution of apparently new (from the perspective of sequence analysis) protein domains and increasing regulatory complexity by domain accretion both quantitatively and qualitatively (recruitment of novel domains with preexisting ones) are two features that we observe in humans. Perhaps the best illustration of this trend is the C2H2 zinc finger–containing transcription factors, where we see expansion in the number of domains per protein, together with vertebrate-specific domains such as KRAB and SCAN. Recent reports on the prominent use of internal ribosomal entry sites in the human genome to regulate translation of specific classes of proteins suggests that this is an area that needs further research to identify the full extent of this process in the human genome (*151*). At the posttranslational level, although we provide examples of expansions of some protein families involved in these modifications, further experimental evidence is required to evaluate whether this is correlated with increased complexity in protein processing. Posttranscriptional processing and the extent of isoform generation in the human remain to be cataloged in their entirety. Given the conserved nature of the spliceosomal machinery, further analysis will be required to dissect regulation at this level.

Conclusions

8.1 The whole-genome sequencing approach versus BAC by BAC

Experience in applying the whole-genome shotgun sequencing approach to a diverse group of organisms with a wide range of genome sizes and repeat content allows us to assess its strengths and weaknesses. With the success of the method for a large number of microbial genomes, *Drosophila*, and now the human, there can be no doubt concerning the utility of this method. The large number of microbial genomes that have been sequenced by this method (*15*, *80*, *152*) demonstrate that megabase-sized genomes can be sequenced efficiently without any input other that the de novo mate-paired sequences. With more complex genomes like those of *Drosophila* or human, map information, in the form of well-ordered markers, has been critical for long-range ordering of scaffolds. For joining scaffolds into chromosomes, the quality of the map (in terms of the order of the markers) is more important than the number of markers per se. Although this mapping could have been performed concurrently with sequenc- the prior existence of mapping data was cial. During the sequencing of the *A. thaliana* genome, sequencing of individual BAC clones permitted extension of the se-

Table 19 (*Continued*)

Panther family/subfamily*	H	F	W	Y	A
Transcription factors/chromatin organization					
C2H2 zinc finger–containing†	607	232	79	28	8
COE	7	1	1	0	0
CREB	7	1	2	0	0
ETS-related	25	8	10	0	0
Forkhead-related	34	19	15	4	0
FOS	8	2	1	0	0
Groucho	13	2	1	0	0
Histone H1	5	0	1	0	0
Histone H2A	24	1	17	3	13
Histone H2B	21	1	17	2	12
Histone H3	28	2	24	2	16
Histone H4	9	1	16	1	8
Homeotic†	168	104	74	4	78
ABD-B	5	0	0	0	0
Bithoraxoid	1	8	1	0	0
Iroquois class	7	3	1	0	0
Distal-less	5	2	1	0	0
Engrailed	2	2	1	0	0
LIM-containing	17	8	3	0	0
MEIS/KNOX class	9	4	4	2	26
NK-3/NK-2 class	9	4	5	0	0
Paired box	38	28	23	0	2
Six	5	3	4	0	0
Leucine zipper	6	0	0	0	0
Nuclear hormone receptor†	59	25	183	1	4
Pou-related	15	5	4	1	0
Runt-related	3	4	2	0	0
ECM adhesion					
Cadherin	113	17	16	0	0
Claudin	20	0	0	0	0
Complement receptor-related	22	8	6	0	0
Connexin	14	0	0	0	0
Galectin	12	5	22	0	0
Glypican	13	2	1	0	0
ICAM	6	0	0	0	0
Integrin alpha	24	7	4	0	1
Integrin beta	9	2	2	0	0
LDL receptor family	26	19	20	0	2
Proteoglycans	22	9	7	0	5
Apoptosis					
Bcl-2	12	1	0	0	0
Calpain	22	4	11	1	3
Calpain inhibitor	4	0	0	0	1
Caspase	13	7	3	0	0
Hemostasis					
ADAM/ADAMTS	51	9	12	0	0
Fibronectin	3	0	0	0	0
Globin	10	2	3	0	3
Matrix metalloprotease	19	2	7	0	3
Serum amyloid A	4	0	0	0	0
Serum amyloid P (subfamily of Pentaxin)	2	0	0	0	0
Serum paraoxonase/arylesterase	4	0	3	0	0
Serum albumin	4	0	0	0	0
Transglutaminase	10	1	0	0	0
Other enzymes					
Cytochrome p450	60	89	83	3	256
GAPDH	46	3	4	3	8
Heparan sulfotransferase	11	4	2	0	0
Splicing and translation					
EF-1alpha	56	13	10	6	13
Ribonucleoproteins†	269	135	104	60	265
Ribosomal proteins†	812	111	80	117	256

*The table lists Panther families or subfamilies relevant to the text that either (i) are not specifically represented by Pfam (Table 18) or (ii) differ in counts from the corresponding Pfam models. †This class represents a number of different families in the same Panther molecular function subcategory. ‡This count includes only rhodopsin-class, secretin-class, and metabotropic glutamate-class GPCRs.

quence well into centromeric regions and allowed high-quality resolution of complex repeat regions. Likewise, in *Drosophila*, the BAC physical map was most useful in regions near the highly repetitive centromeres and telomeres. WGA has been found to deliver excellent-quality reconstructions of the unique regions of the genome. As the genome size, and more importantly the repetitive content, increases, the WGA approach delivers less of the repetitive sequence.

The cost and overall efficiency of clone-by-clone approaches makes them difficult to justify as a stand-alone strategy for future large-scale genome-sequencing projects. Specific applications of BAC-based or other clone mapping and sequencing strategies to resolve ambiguities in sequence assembly that cannot be efficiently resolved with computational approaches alone are clearly worth exploring. Hybrid approaches to whole-genome sequencing will only work if there is sufficient coverage in both the whole-genome shotgun phase and the BAC clone sequencing phase. Our experience with human genome assembly suggests that this will require at least 3× coverage of both whole-genome and BAC shotgun sequence data.

8.2 The low gene number in humans

We have sequenced and assembled ~95% of the euchromatic sequence of *H. sapiens* and used a new automated gene prediction method to produce a preliminary catalog of the human genes. This has provided a major surprise: We have found far fewer genes (26,000 to 38,000) than the earlier molecular predictions (50,000 to over 140,000). Whatever the reasons for this current disparity, only detailed annotation, comparative genomics (particularly using the *Mus musculus* genome), and careful molecular dissection of complex phenotypes will clarify this critical issue of the basic "parts list" of our genome. Certainly, the analysis is still incomplete and considerable refinement will occur in the years to come as the precise structure of each transcription unit is evaluated. A good place to start is to determine why the gene estimates derived from EST data are so discordant with our predictions. It is likely that the following contribute to an inflated gene number derived from ESTs: the variable lengths of 3′- and 5′-untranslated leaders and trailers; the little-understood vagaries of RNA processing that often leave intronic regions in an unspliced condition; the finding that nearly 40% of human genes are alternatively spliced (*153*); and finally, the unsolved technical problems in EST library construction where contamination from heterogeneous nuclear RNA and genomic DNA are not uncommon. Of course, it is possible that there are genes that remain unpredicted owing to the absence of EST or protein data to support them, although our use of mouse genome data for predicting genes should limit this number. As was true at the beginning of genome sequencing, ultimately it will be necessary to measure mRNA in specific cell types to demonstrate the presence of a gene.

J. B. S. Haldane speculated in 1937 that a population of organisms might have to pay a price for the number of genes it can possibly carry. He theorized that when the number of genes becomes too large, each zygote carries so many new deleterious mutations that the population simply cannot maintain itself. On the basis of this premise, and on the basis of available mutation rates and x-ray–induced mutations at specific loci, Muller, in 1967 (*154*), calculated that the mammalian genome would contain a maximum of not much more than 30,000 genes (*155*). An estimate of 30,000 gene loci for humans was also arrived at by Crow and Kimura (*156*). Muller's estimate for *D. melanogaster* was 10,000 genes, compared to 13,000 derived by annotation of the fly genome (*26*, *27*). These arguments for the theoretical maximum gene number were based on simplified ideas of genetic load—that all genes have a certain low rate of mutation to a deleterious state. However, it is clear that many mouse, fly, worm, and yeast knockout mutations lead to almost no discernible phenotypic perturbations.

The modest number of human genes means that we must look elsewhere for the mechanisms that generate the complexities inherent in human development and the sophisticated signaling systems that maintain homeostasis. There are a large number of ways in which the functions of individual genes and gene products are regulated. The degree of "openness" of chromatin structure and hence transcriptional activity is regulated by protein complexes that involve histone and DNA enzymatic modifications. We enumerate many of the proteins that are likely involved in nuclear regulation in Table 19. The location, timing, and quantity of transcription are intimately linked to nuclear signal transduction events as well as by the tissue-specific expression of many of these proteins. Equally important are regulatory DNA elements that include insulators, repeats, and endogenous viruses (*157*); methylation of CpG islands in imprinting (*158*); and promoter-enhancer and intronic regions that modulate transcription. The spliceosomal machinery consists of multisubunit proteins (Table 19) as well as structural and catalytic RNA elements (*159*) that regulate transcript structure through alternative start and termination sites and splicing. Hence, there is a need to study different classes of RNA molecules (*160*) such as small nucleolar RNAs, antisense riboregulator RNA, RNA involved in X-dosage compensation, and other structural RNAs to appreciate their precise role in regulating gene expression. The phenomenon of RNA editing in which coding changes occur directly at the level of mRNA is of clinical and biological relevance (*161*). Finally, examples of translational control include internal ribosomal entry sites that are found in proteins involved in cell cycle regulation and apoptosis (*162*). At the protein level, minor alterations in the nature of protein-protein interactions, protein modifications, and localization can have dramatic effects on cellular physiology (*163*). This dynamic system therefore has many ways to modulate activity, which suggests that definition of complex systems by analysis of single genes is unlikely to be entirely successful.

In situ studies have shown that the human genome is asymmetrically populated with G+C content, CpG islands, and genes (*68*). However, the genes are not distributed quite as unequally as had been predicted (Table 9) (*69*). The most G+C-rich fraction of the genome, H3 isochores, constitute more of the genome than previously thought (about 9%), and are the most gene-dense fraction, but contain only 25% of the genes, rather than the predicted ~40%. The low G+C L isochores make up 65% of the genome, and 48% of the genes. This inhomogeneity, the net result of millions of years of mammalian gene duplication, has been described as the "desertification" of the vertebrate genome (*71*). Why are there clustered regions of high and low gene density, and are these accidents of history or driven by selection and evolution? If these deserts are dispensable, it ought to be possible to find mammalian genomes that are far smaller in size than the human genome. Indeed, many species of bats have genome sizes that are much smaller than that of humans; for example, *Miniopterus*, a species of Italian bat, has a genome size that is only 50% that of humans (*164*). Similarly, *Muntiacus*, a species of Asian barking deer, has a genome size that is ~70% that of humans.

8.3 Human DNA sequence variation and its distribution across the genome

This is the first eukaryotic genome in which a nearly uniform ascertainment of polymorphism has been completed. Although we have identified and mapped more than 3 million SNPs, this by no means implies that the task of finding and cataloging SNPs is complete. These represent only a fraction of the SNPs present in the human population as a whole. Nevertheless, this first glimpse at genome-wide variation has revealed strong inhomogeneities in the distribution of SNPs across the genome. Polymorphism in DNA carries with it a snapshot of the past operation of population genetic forces, including mutation, migration, selection, and genetic drift. The availability of a dense array of SNPs will allow questions related to each of these factors to be addressed on a genome-wide basis. SNP studies can establish the range of haplo-

types present in subjects of different ethnogeographic origins, providing insights into population history and migration patterns. Although such studies have suggested that modern human lineages derive from Africa, many important questions regarding human origins remain unanswered, and more analyses using detailed SNP maps will be needed to settle these controversies. In addition to providing evidence for population expansions, migration, and admixture, SNPs can serve as markers for the extent of evolutionary constraint acting on particular genes. The correlation between patterns of intraspecies and interspecies genetic variation may prove to be especially informative to identify sites of reduced genetic diversity that may mark loci where sequence variations are not tolerated.

The remarkable heterogeneity in SNP density implies that there are a variety of forces acting on polymorphism—sparse regions may have lower SNP density because the mutation rate is lower, because most of those regions have a lower fraction of mutations that are tolerated, or because recent strong selection in favor of a newly arisen allele "swept" the linked variation out of the population (*165*). The effect of random genetic drift also varies widely across the genome. The nonrecombining portion of the Y chromosome faces the strongest pressure from random drift because there are roughly one-quarter as many Y chromosomes in the population as there are autosomal chromosomes, and the level of polymorphism on the Y is correspondingly less. Similarly, the X chromosome has a smaller effective population size than the autosomes, and its nucleotide diversity is also reduced. But even across a single autosome, the effective population size can vary because the density of deleterious mutations may vary. Regions of high density of deleterious mutations will see a greater rate of elimination by selection, and the effective population size will be smaller (*166*). As a result, the density of even completely neutral SNPs will be lower in such regions. There is a large literature on the association between SNP density and local recombination rates in *Drosophila*, and it remains an important task to assess the strength of this association in the human genome, because of its impact on the design of local SNP densities for disease-association studies. It also remains an important task to validate SNPs on a genomic scale in order to assess the degree of heterogeneity among geographic and ethnic populations.

8.4 Genome complexity

We will soon be in a position to move away from the cataloging of individual components of the system, and beyond the simplistic notions of "this binds to that, which then docks on this, and then the complex moves there. . . ." (*167*) to the exciting area of network perturbations, nonlinear responses and thresholds, and their pivotal role in human diseases.

The enumeration of other "parts lists" reveals that in organisms with complex nervous systems, neither gene number, neuron number, nor number of cell types correlates in any meaningful manner with even simplistic measures of structural or behavioral complexity. Nor would they be expected to; this is the realm of nonlinearities and epigenesis (*168*). The 520 million neurons of the common octopus exceed the neuronal number in the brain of a mouse by an order of magnitude. It is apparent from a comparison of genomic data on the mouse and human, and from comparative mammalian neuroanatomy (*169*), that the morphological and behavioral diversity found in mammals is underpinned by a similar gene repertoire and similar neuroanatomies. For example, when one compares a pygmy marmoset (which is only 4 inches tall and weighs about 6 ounces) to a chimpanzee, the brain volume of this minute primate is found to be only about 1.5 cm^3, two orders of magnitude less than that of a chimp and three orders less than that of humans. Yet the neuroanatomies of all three brains are strikingly similar, and the behavioral characteristics of the pygmy marmoset are little different from those of chimpanzees. Between humans and chimpanzees, the gene number, gene structures and functions, chromosomal and genomic organizations, and cell types and neuroanatomies are almost indistinguishable, yet the developmental modifications that predisposed human lineages to cortical expansion and development of the larynx, giving rise to language, culminated in a massive singularity that by even the simplest of criteria made humans more complex in a behavioral sense.

Simple examination of the number of neurons, cell types, or genes or of the genome size does not alone account for the differences in complexity that we observe. Rather, it is the interactions within and among these sets that result in such great variation. In addition, it is possible that there are "special cases" of regulatory gene networks that have a disproportionate effect on the overall system. We have presented several examples of "regulatory genes" that are significantly increased in the human genome compared with the fly and worm. These include extracellular ligands and their cognate receptors (e.g., wnt, frizzled, TGF-β, ephrins, and connexins), as well as nuclear regulators (e.g., the KRAB and homeodomain transcription factor families), where a few proteins control broad developmental processes. The answers to these "complexities" perhaps lie in these expanded gene families and differences in the regulatory control of ancient genes, proteins, pathways, and cells.

8.5 Beyond single components

While few would disagree with the intuitive conclusion that Einstein's brain was more complex than that of *Drosophila*, closer comparisons such as whether the set of predicted human proteins is more complex than the protein set of *Drosophila*, and if so, to what degree, are not straightforward, since protein, protein domain, or protein-protein interaction measures do not capture context-dependent interactions that underpin the dynamics underlying phenotype.

Currently, there are more than 30 different mathematical descriptions of complexity (*170*). However, we have yet to understand the mathematical dependency relating the number of genes with organism complexity. One pragmatic approach to the analysis of biological systems, which are composed of nonidentical elements (proteins, protein complexes, interacting cell types, and interacting neuronal populations), is through graph theory (*171*). The elements of the system can be represented by the vertices of complex topographies, with the edges representing the interactions between them. Examination of large networks reveals that they can self-organize, but more important, they can be particularly robust. This robustness is not due to redundancy, but is a property of inhomogeneously wired networks. The error tolerance of such networks comes with a price; they are vulnerable to the selection or removal of a few nodes that contribute disproportionately to network stability. Gene knockouts provide an illustration. Some knockouts may have minor effects, whereas others have catastrophic effects on the system. In the case of vimentin, a supposedly critical component of the cytoplasmic intermediate filament network of mammals, the knockout of the gene in mice reveals them to be reproductively normal, with no obvious phenotypic effects (*172*), and yet the usually conspicuous vimentin network is completely absent. On the other hand, ~30% of knockouts in *Drosophila* and mice correspond to critical nodes whose reduction in gene product, or total elimination, causes the network to crash most of the time, although even in some of these cases, phenotypic normalcy ensues, given the appropriate genetic background. Thus, there are no "good" genes or "bad" genes, but only networks that exist at various levels and at different connectivities, and at different states of sensitivity to perturbation. Sophisticated mathematical analysis needs to be constantly evaluated against hard biological data sets that specifically address network dynamics. Nowhere is this more critical than in attempts to come to grips with "complexity," particularly because deconvoluting and correcting complex networks that have undergone perturbation, and have resulted in human diseases, is the greatest significant challenge now facing us.

It has been predicted for the last 15 years that complete sequencing of the human ge-

nome would open up new strategies for human biological research and would have a major impact on medicine, and through medicine and public health, on society. Effects on biomedical research are already being felt. This assembly of the human genome sequence is but a first, hesitant step on a long and exciting journey toward understanding the role of the genome in human biology. It has been possible only because of innovations in instrumentation and software that have allowed automation of almost every step of the process from DNA preparation to annotation. The next steps are clear: We must define the complexity that ensues when this relatively modest set of about 30,000 genes is expressed. The sequence provides the framework upon which all the genetics, biochemistry, physiology, and ultimately phenotype depend. It provides the boundaries for scientific inquiry. The sequence is only the first level of understanding of the genome. All genes and their control elements must be identified; their functions, in concert as well as in isolation, defined; their sequence variation worldwide described; and the relation between genome variation and specific phenotypic characteristics determined. Now we know what we have to explain.

Another paramount challenge awaits: public discussion of this information and its potential for improvement of personal health. Many diverse sources of data have shown that any two individuals are more than 99.9% identical in sequence, which means that all the glorious differences among individuals in our species that can be attributed to genes falls in a mere 0.1% of the sequence. There are two fallacies to be avoided: determinism, the idea that all characteristics of the person are "hard-wired" by the genome; and reductionism, the view that with complete knowledge of the human genome sequence, it is only a matter of time before our understanding of gene functions and interactions will provide a complete causal description of human variability. The real challenge of human biology, beyond the task of finding out how genes orchestrate the construction and maintenance of the miraculous mechanism of our bodies, will lie ahead as we seek to explain how our minds have come to organize thoughts sufficiently well to investigate our own existence.

References and Notes

1. R. L. Sinsheimer, *Genomics* 5, 954 (1989); U.S. Department of Energy, Office of Health and Environmental Research, *Sequencing the Human Genome: Summary Report of the Santa Fe Workshop*, Santa Fe, NM, 3 to 4 March 1986 (Los Alamos National Laboratory, Los Alamos, NM, 1986).
2. R. Cook-Deegan, *The Gene Wars: Science, Politics, and the Human Genome* (Norton, New York, 1996).
3. F. Sanger *et al.*, *Nature* 265, 687 (1977).
4. P. H. Seeburg *et al.*, *Trans. Assoc. Am. Physicians* 90, 109 (1977).
5. E. C. Strauss, J. A. Kobori, G. Siu, L. E. Hood, *Anal. Biochem.* 154, 353 (1986).
6. J. Gocayne *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 84, 8296 (1987).
7. A. Martin-Gallardo *et al.*, *DNA Sequence* 3, 237 (1992); W. R. McCombie *et al.*, *Nature Genet.* 1, 348 (1992); M. A. Jensen *et al.*, *DNA Sequence* 1, 233 (1991).
8. M. D. Adams *et al.*, *Science* 252, 1651 (1991).
9. M. D. Adams *et al.*, *Nature* 355, 632 (1992); M. D. Adams, A. R. Kerlavage, C. Fields, J. C. Venter, *Nature Genet.* 4, 256 (1993); M. D. Adams, M. B. Soares, A. R. Kerlavage, C. Fields, J. C. Venter, *Nature Genet.* 4, 373 (1993); M. H. Polymeropoulos *et al.*, *Nature Genet.* 4, 381 (1993); M. Marra *et al.*, *Nature Genet.* 21, 191 (1999).
10. M. D. Adams *et al.*, *Nature* 377, 3 (1995); O. White *et al.*, *Nucleic Acids Res.* 21, 3829 (1993).
11. F. Sanger, A. R. Coulson, G. F. Hong, D. F. Hill, G. B. Petersen, *J. Mol. Biol.* 162, 729 (1982).
12. B. W. J. Mahy, J. J. Esposito, J. C. Venter, *Am. Soc. Microbiol. News* 57, 577 (1991).
13. R. D. Fleischmann *et al.*, *Science* 269, 496 (1995).
14. C. M. Fraser *et al.*, *Science* 270, 397 (1995).
15. C. J. Bult *et al.*, *Science* 273, 1058 (1996); J. F. Tomb *et al.*, *Nature* 388, 539 (1997); H. P. Klenk *et al.*, *Nature* 390, 364 (1997).
16. J. C. Venter, H. O. Smith, L. Hood, *Nature* 381, 364 (1996).
17. H. Schmitt *et al.*, *Genomics* 33, 9 (1996).
18. S. Zhao *et al.*, *Genomics* 63, 321 (2000).
19. X. Lin *et al.*, *Nature* 402, 761 (1999).
20. J. L. Weber, E. W. Myers, *Genome Res.* 7, 401 (1997).
21. P. Green, *Genome Res.* 7, 410 (1997).
22. E. Pennisi, *Science* 280, 1185 (1998).
23. J. C. Venter *et al.*, *Science* 280, 1540 (1998).
24. M. D. Adams *et al.*, *Nature* 368, 474 (1994).
25. E. Marshall, E. Pennisi, *Science* 280, 994 (1998).
26. M. D. Adams *et al.*, *Science* 287, 2185 (2000).
27. G. M. Rubin *et al.*, *Science* 287, 2204 (2000).
28. E. W. Myers *et al.*, *Science* 287, 2196 (2000).
29. F. S. Collins *et al.*, *Science* 282, 682 (1998).
30. International Human Genome Sequencing Consortium (2001), *Nature* 409, 860 (2001).
31. Institutional review board: P. Calabresi (chairman), H. P. Freeman, C. McCarthy, A. L. Caplan, G. D. Rogell, J. Karp, M. K. Evans, B. Margus, C. L. Carter, R. A. Millman, S. Broder.
32. Eligibility criteria for participation in the study were as follows: prospective donors had to be 21 years of age or older, not pregnant, and capable of giving an informed consent. Donors were asked to self-define their ethnic backgrounds. Standard blood bank screens (screening for HIV, hepatitis viruses, and so forth) were performed on all samples at the clinical laboratory prior to DNA extraction in the Celera laboratory. All samples that tested positive for transmissible viruses were ineligible and were discarded. Karyotype analysis was performed on peripheral blood lymphocytes from all samples selected for sequencing; all were normal. A two-staged consent process for prospective donors was employed. The first stage of the consent process provided information about the genome project, procedures, and risks and benefits of participating. The second stage of the consent process involved answering follow-up questions and signing consent forms, and was conducted about 48 hours after the first.
33. DNA was isolated from blood (*173*) or sperm. For sperm, a washed pellet (100 μl) was lysed in a suspension (1 ml) containing 0.1 M NaCl, 10 mM tris-Cl–20 mM EDTA (pH 8), 1% SDS, 1 mg proteinase K, and 10 mM dithiothreitol for 1 hour at 37°C. The lysate was extracted with aqueous phenol and with phenol/chloroform. The DNA was ethanol precipitated and dissolved in 1 ml TE buffer. To make genomic libraries, DNA was randomly sheared, end-polished with consecutive BAL31 nuclease and T4 DNA polymerase treatments, and size-selected by electrophoresis on 1% low-melting-point agarose. After ligation to Bst XI adapters (Invitrogen, catalog no. N408-18), DNA was purified by three rounds of gel electrophoresis to remove excess adapters, and the fragments, now with 3′-CACA overhangs, were inserted into Bst XI-linearized plasmid vector with 3′-TGTG overhangs. Libraries with three different average sizes of inserts were constructed: 2, 10, and 50 kbp. The 2-kbp fragments were cloned in a high-copy pUC18 derivative. The 10- and 50-kbp fragments were cloned in a medium-copy pBR322 derivative. The 2- and 10-kbp libraries yielded uniform-sized large colonies on plating. However, the 50-kbp libraries produced many small colonies and inserts were unstable. To remedy this, the 50-kbp libraries were digested with Bgl II, which does not cleave the vector, but generally cleaved several times within the 50-kbp insert. A 1264-bp Bam HI kanamycin resistance cassette (purified from pUCK4; Amersham Pharmacia, catalog no. 27-4958-01) was added and ligation was carried out at 37°C in the continual presence of Bgl II. As Bgl II–Bgl II ligations occurred, they were continually cleaved, whereas Bam HI–Bgl II ligations were not cleaved. A high yield of internally deleted circular library molecules was obtained in which the residual insert ends were separated by the kanamycin cassette DNA. The internally deleted libraries, when plated on agar containing ampicillin (50 μg/ml), carbenicillin (50 μg/ml), and kanamycin (15 μg/ml), produced relatively uniform large colonies. The resulting clones could be prepared for sequencing using the same procedures as clones from the 10-kbp libraries.
34. Transformed cells were plated on agar diffusion plates prepared with a fresh top layer containing no antibiotic poured on top of a previously set bottom layer containing excess antibiotic, to achieve the correct final concentration. This method of plating permitted the cells to develop antibiotic resistance before being exposed to antibiotic without the potential clone bias that can be introduced through liquid outgrowth protocols. After colonies had grown, QBot (Genetix, UK) automated colony-picking robots were used to pick colonies meeting stringent size and shape criteria and to inoculate 384-well microtiter plates containing liquid growth medium. Liquid cultures were incubated overnight, with shaking, and were scored for growth before passing to template preparation. Template DNA was extracted from liquid bacterial culture using a procedure based upon the alkaline lysis miniprep method (*173*) adapted for high throughput processing in 384-well microtiter plates. Bacterial cells were lysed; cell debris was removed by centrifugation; and plasmid DNA was recovered by isopropanol precipitation and resuspended in 10 mM tris-HCl buffer. Reagent dispensing operations were accomplished using Titertek MAP 8 liquid dispensing systems. Plate-to-plate liquid transfers were performed using Tomtec Quadra 384 Model 320 pipetting robots. All plates were tracked throughout processing by unique plate barcodes. Mated sequencing reads from opposite ends of each clone insert were obtained by preparing two 384-well cycle sequencing reaction plates from each plate of plasmid template DNA using ABI-PRISM BigDye Terminator chemistry (Applied Biosystems) and standard M13 forward and reverse primers. Sequencing reactions were prepared using the Tomtec Quadra 384-320 pipetting robot. Parent-child plate relationships and, by extension, forward-reverse sequence mate pairs were established by automated plate barcode reading by the onboard barcode reader and were recorded by direct LIMS communication. Sequencing reaction products were purified by alcohol precipitation and were dried, sealed, and stored at 4°C in the dark until needed for sequencing, at which time the reaction products were resuspended in deionized formamide and sealed immediately to prevent degradation. All sequence data were generated using a single sequencing platform, the ABI PRISM 3700 DNA Analyzer. Sample sheets were created at load time using a Java-based application that facilitates barcode scanning of the sequencing plate barcode, retrieves sample information from the central LIMS, and reserves unique trace identifiers. The application permitted a single sample sheet file in the linking directory and deleted previously created sample sheet files immediately upon scanning of a

sample plate barcode, thus enhancing sample sheet-to-plate associations.
35. F. Sanger, S. Nicklen, A. R. Coulson, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463 (1977); J. M. Prober *et al.*, *Science* **238**, 336 (1987).
36. Celera's computing environment is based on Compaq Computer Corporation's Alpha system technology running the Tru64 Unix operating system. Celera uses these Alphas as Data Servers and as nodes in a Virtual Compute Farm, all of which are connected to a fully switched network operating at Fast Ethernet speed (for the VCF) and gigabit Ethernet speed (for data servers). Load balancing and scheduling software manages the submission and execution of jobs, based on central processing unit (CPU) speed, memory requirements, and priority. The Virtual Compute Farm is composed of 440 Alpha CPUs, which includes model EV6 running at a clock speed of 400 MHz and EV67 running at 667 MHz. Available memory on these systems ranges from 2 GB to 8 GB. The VCF is used to manage trace file processing, and annotation. Genome assembly was performed on a GS 160 running 16 EV67s (667 MHz) and 64 GB of memory, and 10 ES40s running 4 EV6s (500 MHz) and 32 GB of memory. A total of 100 terabytes of physical disk storage was included in a Storage Area Network that was available to systems across the environment. To ensure high availability, file and database servers were configured as 4-node Alpha TruClusters, so that services would fail over in the event of hardware or software failure. Data availability was further enhanced by using hardware- and software-based disk mirroring (RAID-0), disk striping (RAID-1), and disk striping with parity (RAID-5).
37. Trace processing generates quality values for base calls by means of Paracel's TraceTuner, trims sequence reads according to quality values, trims vector and adapter sequence from high-quality reads, and screens sequences for contaminants. Similar in design and algorithm to the phred program (*174*), TraceTuner reports quality values that reflect the log-odds score of each base being correct. Read quality was evaluated in 50-bp windows, each read being trimmed to include only those consecutive 50-bp segments with a minimum mean accuracy of 97%. End windows (both ends of the trace) of 1, 5, 10, 25, and 50 bases were trimmed to a minimum mean accuracy of 98%. Every read was further checked for vector and contaminant matches of 50 bp or more, and if found, the read was removed from consideration. Finally, any match to the 5′ vector splice junction in the initial part of a read was removed.
38. National Center for Biotechnology Information (NCBI); available at www.ncbi.nlm.nih.gov/.
39. NCBI; available at www.ncbi.nlm.nih.gov/HTGS/.
40. All bactigs over 3 kbp were examined for coverage by Celera mate pairs. An interval of a bactig was deemed an assembly error where there were no mate pairs spanning the interval and at least two reads that should have their mate on the other side of the interval but did not. In other words, there was no mate pair evidence supporting a join in the breakpoint interval and at least two mate pairs contradicting the join. By this criterion, we detected and broke apart bactigs at 13,037 locations, or equivalently, we found 2.13% of the bactigs to be misassembled.
41. We considered a BAC entry to be chimeric if, by the Lander-Waterman statistic (*175*), the odds were 0.99 or more that the assembly we produced was inconsistent with the sequence coming from a single source. By this criterion, 714 or 2.2% of BAC entries were deemed chimeric.
42. G. Myers, S. Selznick, Z. Zhang, W. Miller, *J. Comput. Biol.* **3**, 563 (1996).
43. E. W. Myers, J. L. Weber, in *Computational Methods in Genome Research*, S. Suhai, Ed. (Plenum, New York, 1996), pp. 73–89.
44. P. Deloukas *et al.*, *Science* **282**, 744 (1998).
45. M. A. Marra *et al.*, *Genome Res.* **7**, 1072 (1997).
46. Zhang *et al.*, data not shown.
47. Shredded bactigs were located on long CSA scaffolds (>500 kbp) and the distribution of these fragments on the scaffolds was analyzed. If the spread of these fragments was greater than four times the reported BAC length, the BAC was considered to be chimeric. In addition, if >20% of bactigs of a given BAC were found on a different scaffolds that were not adjacent in map position, then the BAC was also considered as chimeric. The total chimeric BACs divided by the number of BACs used for CSA gave the minimal estimate of chimerism rate.
48. M. Hattori *et al.*, *Nature* **405**, 311 (2000).
49. I. Dunham *et al.*, *Nature* **402**, 489 (1999).
50. A. B. Carvalho, B. P. Lazzaro, A. G. Clark, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 13239 (2000).
51. The International RH Mapping Consortium, available at www.ncbi.nlm.nih.gov/genemap99/.
52. See http://ftp.genome.washington.edu/RM/RepeatMasker.html.
53. G. D. Schuler, *Trends Biotechnol.* **16**, 456 (1998).
54. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990).
55a. M. Olivier *et al.*, *Science* **291**, 1298 (2001).
55b. See http://genome.ucsc.edu/.
56. N. Chaudhari, W. E. Hahn, *Science* **220**, 924 (1983); R. J. Milner, J. G. Sutcliffe, *Nucleic Acids Res.* **11**, 5497 (1983).
57. D. Dickson, *Nature* **401**, 311 (1999).
58. B. Ewing, P. Green, *Nature Genet.* **25**, 232 (2000).
59. H. Roest Crollius *et al.*, *Nature Genet.* **25**, 235 (2000).
60. M. Yandell, in preparation.
61. K. D. Pruitt, K. S. Katz, H. Sicotte, D. R. Maglott, *Trends Genet.* **16**, 44 (2000).
62. Scaffolds containing greater than 10 kbp of sequence were analyzed for features of biological importance through a series of computational steps, and the results were stored in a relational database. For scaffolds greater than one megabase, the sequence was cut into single megabase pieces before computational analysis. All sequence was masked for complex repeats using Repeatmasker (*52*) before gene finding or homology-based analysis. The computational pipeline required ~7 hours of CPU time per megabase, including repeat masking, or a total compute time of about 20,000 CPU hours. Protein searches were performed against the nonredundant protein database available at the NCBI. Nucleotide searches were performed against human, mouse, and rat Celera Gene Indices (assemblies of cDNA and EST sequences), mouse genomic DNA reads generated at Celera (3×), the Ensembl gene database available at the European Bioinformatics Institute (EBI), human and rodent (mouse and rat) EST data sets parsed from the dbEST database (NCBI), and a curated subset of the RefSeq experimental mRNA database (NCBI). Initial searches were performed on repeat-masked sequence with BLAST 2.0 (*54*) optimized for the Compaq Alpha computeserver and an effective database size of 3×10^9 for BLASTN searches and 1×10^9 for BLASTX searches. Additional processing of each query-subject pair was performed to improve the alignments. All protein BLAST results having an expectation score of $<1 \times 10^{-4}$, human nucleotide BLAST results having an expectation score of $<1 \times 10^{-8}$ with >94% identity, and rodent nucleotide BLAST results having an expectation score of $<1 \times 10^{8}$ with >80% identity were then examined on the basis of their high-scoring pair (HSP) coordinates on the scaffold to remove redundant hits, retaining hits that supported possible alternative splicing. For BLASTX searches, analysis was performed separately for selected model organisms (yeast, mouse, human, *C. elegans*, and *D. melanogaster*) so as not to exclude HSPs from these organisms that support the same gene structure. Sequences producing BLAST hits judged to be informative, nonredundant, and sufficiently similar to the scaffold sequence were then realigned to the genomic sequence with Sim4 for ESTs, and with Lap for proteins. Because both of these algorithms take splicing into account, the resulting alignments usually give a better representation of intron-exon boundaries than standard BLAST analyses and thus facilitate further annotation (both machine and human). In addition to the homology-based analysis described above, three ab initio gene prediction programs were used (*63*).
63. E. C. Uberbacher, Y. Xu, R. J. Mural, *Methods Enzymol.* **266**, 259 (1996); C. Burge, S. Karlin, *J. Mol. Biol.* **268**, 78 (1997); R. J. Mural, *Methods Enzymol.* **303**, 77 (1999); A. A. Salamov, V. V. Solovyev, *Genome Res.* **10**, 516 (2000); Floreal *et al.*, *Genome Res.* **8**, 967 (1998).
64. G. L. Miklos, B. John, *Am. J. Hum. Genet.* **31**, 264 (1979); U. Francke, *Cytogenet. Cell Genet.* **65**, 206 (1994).
65. P. E. Warburton, H. F. Willard, in *Human Genome Evolution*, M. S. Jackson, T. Strachan, G. Dover, Eds. (BIOS Scientific, Oxford, 1996), pp. 121–145.
66. J. E. Horvath, S. Schwartz, E. E. Eichler, *Genome Res.* **10**, 839 (2000).
67. W. A. Bickmore, A. T. Sumner, *Trends Genet.* **5**, 144 (1989).
68. G. P. Holmquist, *Am. J. Hum. Genet.* **51**, 17 (1992).
69. G. Bernardi, *Gene* **241**, 3 (2000).
70. S. Zoubak, O. Clay, G. Bernardi, *Gene* **174**, 95 (1996).
71. S. Ohno, *Trends Genet.* **1**, 160 (1985).
72. K. W. Broman, J. C. Murray, V. C. Sheffield, R. L. White, J. L. Weber, *Am. J. Hum. Genet.* **63**, 861 (1998).
73. M. J. McEachern, A. Krauskopf, E. H. Blackburn, *Annu. Rev. Genet.* **34**, 331 (2000).
74. A. Bird, *Trends Genet.* **3**, 342 (1987).
75. M. Gardiner-Garden, M. Frommer, *J. Mol. Biol.* **196**, 261 (1987).
76. F. Larsen, G. Gundersen, R. Lopez, H. Prydz, *Genomics* **13**, 1095 (1992).
77. S. H. Cross, A. Bird, *Curr. Opin. Genet. Dev.* **5**, 309 (1995).
78. J. Peters, *Genome Biol.* **1**, reviews1028.1 (2000) (http://genomebiology.com/2000/1/5/reviews/1028).
79. C. Grunau, W. Hindermann, A. Rosenthal, *Hum. Mol. Genet.* **9**, 2651 (2000).
80. F. Antequera, A. Bird, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 11995 (1993).
81. S. H. Cross *et al.*, *Mamm. Genome* **11**, 373 (2000).
82. D. Slavov *et al.*, *Gene* **247**, 215 (2000).
83. A. F. Smit, A. D. Riggs, *Nucleic Acids Res.* **23**, 98 (1995).
84. D. J. Elliott *et al.*, *Hum. Mol. Genet.* **9**, 2117 (2000).
85. A. V. Makeyev, A. N. Chkheidze, S. A. Lievhaber, *J. Biol. Chem.* **274**, 24849 (1999).
86. Y. Pan, W. K. Decker, A. H. H. M. Huq, W. J. Craigen, *Genomics* **59**, 282 (1999).
87. P. Nouvel, *Genetica* **93**, 191 (1994).
88. I. Goncalves, L. Duret, D. Mouchiroud, *Genome Res.* **10**, 672 (2000).
89. Lek first compares all proteins in the proteome to one another. Next, the resulting BLAST reports are parsed, and a graph is created wherein each protein constitutes a node; any hit between two proteins with an expectation beneath a user-specified threshold constitutes an edge. Lek then uses this graph to compute a similarity between each protein pair *ij* in the context of the graph as a whole by simply dividing the number of BLAST hits shared in common between the two proteins by the total number of proteins hit by *I* and *J*. This simple metric has several interesting properties. First, because the similarity metric takes into account both the similarity and the differences between the two sequences at the level of BLAST hits, the metric respects the multidomain nature of protein space. Two multidomain proteins, for instance, each containing domains A and B, will have a greater pairwise similarity to each other than either one will have to a protein containing only A or B domains, so long as A-B–containing multidomain proteins are less frequent in the proteome than are single-domain proteins containing A or B domains. A second interesting property of this similarity metric is that it can be used to produce a similarity matrix for the proteome as a whole without having to first produce a multiple alignment for each protein family, an error-prone and very time-consuming process. Finally, the metric does not require that either sequence have significant homology to the other in order to have a defined similarity to each other, only that they

share at least one significant BLAST hit in common. This is an especially interesting property of the metric, because it allows the rapid recovery of protein families from the proteome for which no multiple alignment is possible, thus providing a computational basis for the extension of protein homology searches beyond those of current HMM- and profile-based search methods. Once the whole-proteome similarity matrix has been calculated, Lek first partitions the proteome into single-linkage clusters (*27*) on the basis of one or more shared BLAST hits between two sequences. Next, these single-linkage clusters are further partitioned into subclusters, each member of which shares a user-specified pairwise similarity with the other members of the cluster, as described above. For the purposes of this publication, we have focused on the analysis of single-linkage clusters and what we have termed "complete clusters," e.g., those subclusters for which every member has a similarity metric of 1 to every other member of the subcluster. We believe that the single-linkage and complete clusters are of special interest, in part, because they allow us to estimate and to compare sizes of core protein sets in a rigorous manner. The rationale for this is as follows: if one imagines for a moment a perfect clustering algorithm capable of perfectly partitioning one or more perfectly annotated protein sets into protein families, it is reasonable to assume that the number of clusters will always be greater than, or equal to, the number of single-linkage clusters, because single-linkage clustering is a maximally agglomerative clustering method. Thus, if there exists a single protein in the predicted protein set containing domains A and B, then it will be clustered by single linkage together with all single-domain proteins containing domains A or B. Likewise, for a predicted protein set containing a single multidomain protein, the number of real clusters must always be less than or equal to the number of complete clusters, because it is impossible to place a unique multidomain protein into a complete cluster. Thus, the single-linkage and complete clusters plus singletons should comprise a lower and upper bound of sizes of core protein sets, respectively, allowing us to compare the relative size and complexity of different organisms' predicted protein set.

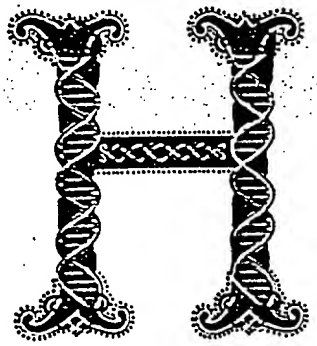
90. T. F. Smith, M. S. Waterman, *J. Mol. Biol.* **147**, 195 (1981).
91. A. L. Delcher *et al.*, *Nucleic Acids Res.* **27**, 2369 (1999).
92. *Arabidopsis* Genome Initiative, *Nature* **408**, 796 (2000).
93. The probability that a contiguous set of proteins is the result of a segmental duplication can be estimated approximately as follows. Given that protein A and B occur on one chromosome, and that A′ and B′ (paralogs of A and B) also exist in the genome, the probability that B′ occurs immediately after A′ is $1/N$, where N is the number of proteins in the set (for this analysis, $N = 26{,}588$). Allowing for B′ to occur as any of the next J-1 proteins [leaving a gap between A′ and B′ increases the probability to $(J - 1)/N$; allowing B′A′ or A′B′ gives a probability of $2(J - 1)/N$]. Considering three genes ABC, the probability of observing A′B′C′ elsewhere in the genome, given that the paralogs exist, is $1/N^2$. Three proteins can occur across a spread of five positions in six ways; more generally, we compute the number of ways that K proteins can be spread across J positions by counting all possible arrangements of $K - 2$ proteins in the $J - 2$ positions between the first and last protein. Allowing for a spread to vary from K positions (no gaps) to J gives
$$L = \sum_{x=K-2}^{J-2} \binom{x}{K-2}$$
arrangements. Thus, the probability of chance occurrence is L/N^{K-1}. Allowing for both sets of genes (e.g., ABC and A′B′C′) to be spread across J positions increases this to L^2/N^{K-1}. The duplicated segment might be rearranged by the operations of reversal or translocation; allowing for M such rearrangements gives us a probability $P = L^2M/N^{K-1}$. For example, the probability of observing a duplicated set of three genes in two different locations, where the three genes occur across a spread of five positions in both locations, is $36/N^2$; the expected number of such matched sets in the predicted protein set is approximately $(N)36/N^2 = 36/N$, a value $\ll 1$. Therefore, any such duplications of three genes are unlikely to result from random rearrangements of the genome. If any of the genes occur in more than two copies, the probability that the apparent duplication has occurred by chance increases. The algorithm for selecting candidate duplications only generates matched protein sets with $P \ll 1$.
94. B. J. Trask *et al.*, *Hum. Mol. Genet.* **7**, 13 (1998); D. Sharon *et al.*, *Genomics* **61**, 24 (1999).
95. W. B. Barbazuk *et al.*, *Genome Res.* **10**, 1351 (2000); A. McLysaght, A. J. Enright, L. Skrabanek, K. H. Wolfe, *Yeast* **17**, 22 (2000); D. W. Burt *et al.*, *Nature* **402**, 411 (1999).
96. Reviewed in L. Skrabanek, K. H. Wolfe, *Curr. Opin. Genet. Dev.* **8**, 694 (1998).
97. P. Taillon-Miller, Z. Gu, Q. Li, L. Hillier, P. Y. Kwok, *Genome Res.* **8**, 748 (1998); P. Taillon-Miller, E. E. Piernot, P. Y. Kwok, *Genome Res.* **9**, 499 (1999).
98. D. Altshuler *et al.*, *Nature* **407**, 513 (2000).
99. G. T. Marth *et al.*, *Nature Genet.* **23**, 452 (1999).
100. W.-H. Li, *Molecular Evolution* (Sinauer, Sunderland, MA, 1997).
101. M. Cargill *et al.*, *Nature Genet.* **22**, 231 (1999).
102. M. K. Halushka *et al.*, *Nature Genet.* **22**, 239 (1999).
103. J. Zhang, T. L. Madden, *Genome Res.* **7**, 649 (1997).
104. M. Nei, *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York, 1987).
105. From the observed coverage of the sequences at each site for each individual, we calculated the probability that a SNP would be detected at the site if it were present. For each level of coverage, there is a binomial sampling of the two homologs for each individual, and a heterozygous site could only be ascertained if both homologs are present, or if two alleles from different individuals are present. With coverage x from a given individual, both homologs are present in the assembly with probability $1 - (1/2)x^{-1}$. Even if both homologs are present, the probability that a SNP is detected is <1 because a fraction of sites failed the quality criteria. Integrating over coverage levels, the binomial sampling, and the quality distribution, we derived an expected number of sites in the genome that were ascertained for polymorphism for each individual. The nucleotide diversity was then the observed number of variable sites divided by the expected number of sites ascertained.
106. M. W. Nachman, V. L. Bauer, S. L. Crowell, C. F. Aquadro, *Genetics* **150**, 1133 (1998).
107. D. A. Nickerson *et al.*, *Nature Genet.* **19**, 233 (1998); D. A. Nickerson *et al.*, *Genomic Res.* **10**, 1532 (2000); L. Jorde *et al.*, *Am. J. Hum. Genet.* **66**, 979 (2000); D. G. Wang *et al.*, *Science* **280**, 1077 (1998).
108. M. Przeworski, R. R. Hudson, A. Di Rienzo, *Trends Genet.* **16**, 296 (2000).
109. S. Tavare, *Theor. Popul. Biol.* **26**, 119 (1984).
110. R. R. Hudson, in *Oxford Surveys in Evolutionary Biology*, D. J. Futuyma, J. D. Antonovics, Eds. (Oxford Univ. Press, Oxford, 1990), vol. 7, pp. 1–44.
111. A. G. Clark *et al.*, *Am. J. Hum. Genet.* **63**, 595 (1998).
112. M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, 1983).
113. H. Kaessmann, F. Heissig, A. von Haeseler, S. Paabo, *Nature Genet.* **22**, 78 (1999).
114. E. L. Sonnhammer, S. R. Eddy, R. Durbin, *Proteins* **28**, 405 (1997).
115. A. Bateman *et al.*, *Nucleic Acids Res.* **28**, 263 (2000).
116. Brief description of the methods used to build the Panther classification. First, the June 2000 release of the GenBank NR protein database (excluding sequences annotated as fragments or mutants) was partitioned into clusters using BLASTP. For the clustering, a seed sequence was randomly chosen, and the cluster was defined as all sequences matching the seed to statistical significance (E-value $< 10^{-5}$) and "globally" alignable (the length of the match region must be >70% and <130% of the length of the seed). If the cluster had more than five members, and at least one from a multicellular eukaryote, the cluster was extended. For the extension step, a hidden Markov Model (HMM) was trained for the cluster, using the SAM software package, version 2. The HMM was then scored against GenBank NR (excluding mutants but including fragments for this step), and all sequences scoring better than a specific (NLL-NULL) score were added to the cluster. The HMM was then retrained (with fixed model length) and all sequences in the cluster were aligned to the HMM to produce a multiple sequence alignment. This alignment was assessed by a number of quality measures. If the alignment failed the quality check, the initial cluster was rebuilt around the seed using a more restrictive E-value, followed by extension, alignment, and reassessment. This process was repeated until the alignment quality was good. The multiple alignment and "general" (i.e., describing the entire cluster, or "family") HMM (*176*) were then used as input into the BETE program (*177*). BETE calculates a phylogenetic tree for the sequences in the alignment. Functional information about the sequences in each cluster were parsed from SwissProt (*178*) and GenBank records. "Tree-attribute viewer" software was used by biologist curators to correlate the phylogenetic tree with protein function. Subfamilies were manually defined on the basis of shared function across subtrees, and were named accordingly. HMMs were then built for each subfamily, using information from both the subfamily and family (K. Sjölander, in preparation). Families were also manually named according to the functions contained within them. Finally, all of the families and subfamilies were classified into categories and subcategories based on their molecular functions. The categorization was done by manual review of the family and subfamily names, by examining SwissProt and GenBank records, and by review of the literature as well as resources on the World Wide Web. The current version (2.0) of the Panther molecular function schema has four levels: category, subcategory, family, and subfamily. Protein sequences for whole eukaryotic genomes (for the predicted human proteins and annotated proteins for fly, worm, yeast, and *Arabidopsis*) were scored against the Panther library of family and subfamily HMMs. If the score was significant (the NLL-NULL score cutoff depends on the protein family), the protein was assigned to the family or subfamily function with the most significant score.
117. C. P. Ponting, J. Schultz, F. Milpetz, P. Bork, *Nucleic Acids Res.* **27**, 229 (1999).
118. A. Goffeau *et al.*, *Science* **274**, 546, 563 (1996).
119. *C. elegans* Sequencing Consortium, *Science* **282**, 2012 (1998).
120. S. A. Chervitz *et al.*, *Science* **282**, 2022 (1998).
121. E. R. Kandel, J. H. Schwartz, T. Jessell, *Principles of Neural Science* (McGraw-Hill, New York, ed. 4, 2000).
122. D. A. Goodenough, J. A. Goliger, D. L. Paul, *Annu. Rev. Biochem.* **65**, 475 (1996).
123. D. G. Wilkinson, *Int. Rev. Cytol.* **196**, 177 (2000).
124. F. Nakamura, R. G. Kalb, S. M. Strittmatter, *J. Neurobiol.* **44**, 219 (2000).
125. P. J. Horner, F. H. Gage, *Nature* **407**, 963 (2000); P. Casaccia-Bonnefil, C. Gu, M. V. Chao, *Adv. Exp. Med. Biol.* **468**, 275 (1999).
126. S. Wang, B. A. Barres, *Neuron* **27**, 197 (2000).
127. M. Geppert, T. C. Sudhof, *Annu. Rev. Neurosci.* **21**, 75 (1998); J. T. Littleton, H. J. Bellen, *Trends Neurosci.* **18**, 177 (1995).
128. A. Maximov, T. C. Sudhof, I. Bezprozvanny, *J. Biol. Chem.* **274**, 24453 (1999).
129. B. Sampo *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 3666 (2000).
130. G. Lemke, *Glia* **7**, 263 (1993).
131. M. Bernfield *et al.*, *Annu. Rev. Biochem.* **68**, 729 (1999).
132. N. Perrimon, M. Bernfield, *Nature* **404**, 725 (2000).
133. U. Lindahl, M. Kusche-Gullberg, L. Kjellen, *J. Biol. Chem.* **273**, 24979 (1998).
134. J. L. Riechmann *et al.*, *Science* **290**, 2105 (2000).
135. T. L. Hurskainen, S. Hirohata, M. F. Seldin, S. S. Apte, *J. Biol. Chem.* **274**, 25555 (1999).

136. R. A. Black, J. M. White, *Curr. Opin. Cell Biol.* **10**, 654 (1998).
137. L. Aravind, V. M. Dixit, E. V. Koonin, *Trends Biochem. Sci.* **24**, 47 (1999).
138. A. G. Uren *et al.*, *Mol. Cell* **6**, 961 (2000).
139. P. Garcia-Meunier, M. Etienne-Julan, P. Fort, M. Piechaczyk, F. Bonhomme, *Mamm. Genome* **4**, 695 (1993).
140. K. Meyer-Siegler *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 8460 (1991).
141. N. R. Mansur, K. Meyer-Siegler, J. C. Wurzer, M. A. Sirover, *Nucleic Acids Res.* **21**, 993 (1993).
142. N. A. Tatton, *Exp. Neurol.* **166**, 29 (2000).
143. N. Kenmochi *et al.*, *Genome Res.* **8**, 509 (1998).
144. F. W. Chen, Y. A. Ioannou, *Int. Rev. Immunol.* **18**, 429 (1999).
145. H. O. Madsen, K. Poulsen, O. Dahl, B. F. Clark, J. P. Hjorth, *Nucleic Acids Res.* **18**, 1513 (1990).
146. D. M. Chambers, J. Peters, C. M. Abbott, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 4463 (1998); A. Khalyfa, B. M. Carlson, J. A. Carlson, E. Wang, *Dev. Dyn.* **216**, 267 (1999).
147. D. Aeschlimann, V. Thomazy, *Connect. Tissue Res.* **41**, 1 (2000).
148. P. Munroe *et al.*, *Nature Genet.* **21**, 142 (1999); S. M. Wu, W. F. Cheung, D. Frazier, D. W. Stafford, *Science* **254**, 1634 (1991); B. Furie *et al.*, *Blood* **93**, 1798 (1999).
149. J. W. Kehoe, C. R. Bertozzi, *Chem. Biol.* **7**, R57 (2000).
150. T. Pawson, P. Nash, *Genes Dev.* **14**, 1027 (2000).
151. A. W. van der Velden, A. A. Thomas, *Int. J. Biochem. Cell Biol.* **31**, 87 (1999).
152. C. M. Fraser *et al.*, *Science* **281**, 375 (1998); H. Tettelin *et al.*, *Science* **287**, 1809 (2000).
153. D. Brett *et al.*, *FEBS Lett.* **474**, 83 (2000).
154. H. J. Muller, H. Kern, *Z. Naturforsch. B* **22**, 1330 (1967).
155. H. J. Muller, in *Heritage from Mendel*, R. A. Brink, Ed. (Univ. of Wisconsin Press, Madison, WI, 1967), p. 419.
156. J. F. Crow, M. Kimura, *Introduction to Population Genetics Theory* (Harper & Row, New York, 1970).
157. K. Kobayashi *et al.*, *Nature* **394**, 388 (1998).
158. A. P. Feinberg, *Curr. Top. Microbiol. Immunol.* **249**, 87 (2000).
159. C. A. Collins, C. Guthrie, *Nature Struct. Biol.* **7**, 850 (2000).
160. S. R. Eddy, *Curr. Opin. Genet. Dev.* **9**, 695 (1999).
161. Q. Wang, J. Khillan, P. Gadue, K. Nishikura, *Science* **290**, 1765 (2000).
162. M. Holcik, N. Sonenberg, R. G. Korneluk, *Trends Genet.* **16**, 469 (2000).
163. T. A. McKinsey, C. L. Zhang, J. Lu, E. N. Olson, *Nature* **408**, 106 (2000).
164. E. Capanna, M. G. M. Romanini, *Caryologia* **24**, 471 (1971).
165. J. Maynard Smith, *J. Theor. Biol.* **128**, 247 (1987).
166. D. Charlesworth, B. Charlesworth, M. T. Morgan, *Genetics* **141**, 1619 (1995).
167. J. E. Bailey, *Nature Biotechnol.* **17**, 616 (1999).
168. R. Maleszka, H. G. de Couet, G. L. Miklos, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 3731 (1998).
169. G. L. Miklos, *J. Neurobiol.* **24**, 842 (1993).
170. J. P. Crutchfield, K. Young, *Phys. Rev. Lett.* **63**, 105 (1989); M. Gell-Mann, S. Lloyd, *Complexity* **2**, 44 (1996).
171. A. L. Barabasi, R. Albert, *Science* **286**, 509 (1999).
172. E. Colucci-Guyon *et al.*, *Cell* **79**, 679 (1994).
173. J. Sambrook, E. F. Fritch, T. Maniatis, *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, ed. 2, 1989).
174. B. Ewing, P. Green, *Genome Res.* **8**, 186 (1998); B. Ewing, L. Hillier, M. C. Wendl, P. Green, *Genome Res.* **8**, 175 (1998).
175. E. S. Lander, M. S. Waterman, *Genomics* **2**, 231 (1988).
176. A. Krogh, K. Sjölander, *J. Mol. Biol.* **235**, 1501 (1994).
177. K. Sjölander, *Proc. Int. Soc. Mol. Biol.* **6**, 165 (1998).
178. A. Bairoch, R. Apweiler, *Nucleic Acids Res.* **28**, 45 (2000).
179. GO, available at www.geneontology.org/.
180. R. L. Tatusov, M. Y. Galperin, D. A. Natale, E. V. Koonin, *Nucleic Acids Res.* **28**, 33 (2000).

181. We thank E. Eichler and J. L. Goldstein for many helpful discussions and critical reading of the manuscript, and A. Caplan for advice and encouragement. We also thank T. Hein, D. Lucas, G. Edwards, and the Celera IT staff for outstanding computational support. The cost of this project was underwritten by the Celera Genomics Group of the Applera Corporation. We thank the Board of Directors of Applera Corporation: J. F. Abely Jr. (retired), R. H. Ayers, J.-L. Bélingard, R. H. Hayes, A. J. Levine, T. E. Martin, C. W. Slayman, O. R. Smith, G. C. St. Laurent Jr., and J. R. Tobin for their vision, enthusiasm, and unwavering support and T. L. White for leadership and advice. Data availability: The genome sequence and additional supporting information are available to academic scientists at the Web site (www.celera.com). Instructions for obtaining a DVD of the genome sequence can be obtained through the Web site. For commercial scientists wishing to verify the results presented here, the genome data are available upon signing a Material Transfer Agreement, which can also be found on the Web site.

5 December 2000; accepted 19 January 2001

THE HUMAN GENOME



Humanity has been given a great gift. With the completion of the human genome sequence, we have received a powerful tool for unlocking the secrets of our genetic heritage and for finding our place among the other participants in the adventure of life.

This week's issue of *Science* contains the report of the sequencing of the human genome from a group of authors led by Craig Venter of Celera Genomics. The report of the sequencing of the human genome from the publicly funded consortium of laboratories led by Francis Collins appears in this week's *Nature*. This stunning achievement has been portrayed—often unfairly—as a competition between two ventures, one public and one private. That characterization detracts from the awesome accomplishment jointly unveiled this week. In truth, each project contributed to the other. The inspired vision that launched the publicly funded project roughly 10 years ago reflected, and now rewards, the confidence of those who believe that the pursuit of large-scale fundamental problems in the life sciences is in the national interest. The technical innovation and drive of Craig Venter and his colleagues made it possible to celebrate this accomplishment far sooner than was believed possible. Thus, we can salute what has become, in the end, not a contest but a marriage (perhaps encouraged by shotgun) between public funding and private entrepreneurship.

A historic moment for the scientific endeavor.

There are excellent scientific reasons for applauding an outcome that has given us two winners. Two sequences are better than one; the opportunity for comparison and convergence is invaluable. Indeed, a real-world proof of the importance of access to both sets of data can be found in the pages of this issue of *Science*, in the comparative analysis by Olivier *et al.* (p. 1298).

Although we have made the point before, it is worth repeating that the sequencing of the human genome represents, not an ending, but the beginning of a new approach to biology. As Galas says in his Viewpoint (p. 1257), the knowledge that all of the genetic components of any process can be identified will give extraordinary new power to scientists. Because of this breakthrough, research can evolve from analyzing the effects of individual genes to a more integrated view that examines whole ensembles of genes as they interact to form a living human being. Several articles in this issue highlight how this approach is already beginning to revolutionize the way we look at human disease.

This has been a massive project, on a scale unparalleled in the history of biology, but of course it has built on the scientific insights of centuries of investigators. By coincidence, this landmark announcement falls during the week of the anniversary of the birth of Charles Darwin. Darwin's message that the survival of a species can depend on its ability to evolve in the face of change is peculiarly pertinent to discussions that have gone on in the past year over access to the Celera data. (Full information regarding the agreements that were reached to make the data available can be found at www.sciencemag.org/feature/data/announcement/gsp.shl.) We are willing to be flexible in allowing data repositories other than the traditional GenBank, while insisting on access to all the data needed to verify conclusions. In this domain, change is everywhere: Commercial researchers are producing more and more potentially valuable sequences, yet (at least in the United States) laws governing databases provide scant protection against piracy. Had the Celera data been kept secret, it would have been a serious loss to the scientific community. We hope that our adaptability in the face of change will enable other proprietary data to be published after peer review, in a way that satisfies our continuing commitment to full access.

It should be no surprise that an achievement so stunning, and so carefully watched, has created new challenges for the scientific venture. *Science* is proud to have played a role in bringing this discovery onto the public stage. It is literally true that this is a historic moment for the scientific endeavor. The human genome has been called the Book of Life. Rather, it is a library, in which, with rules that encourage exploration and reward creativity, we can find many of the books that will help define us and our place in the great tapestry of life.

Barbara R. Jasny and Donald Kennedy

Home **Paracel BLAST Results** Help

MEGABLAST 1.2.3-Paracel [2001-11-20]

Reference:
Zheng Zhang, Scott Schwartz, Lukas Wagner, and Webb Miller (2000),
"A greedy algorithm for aligning DNA sequences",
J Comput Biol 2000; 7(1-2):203-14.
Database: Homo_sapiens.latestgp.fa
26,679 sequences; 200,800,637,119 total letters

Query= 1
(2629 letters)

Sequences producing significant alignments:	Score (bits)	E Value
AC006208.3.1.123943	940	0.0
AC000063.1.1.34478	72	2e-09
AC079799.7.1.172495	54	5e-04

>AC006208.3.1.123943
Length = 123943

Score = 940 bits (474), Expect = 0.0
Identities = 474/474 (100%)
Strand = Plus / Minus

```
Query: 2156  caggtgaagacggacgagcgagtcttgcacacggagcgggggctgctgttccgcaggctt 2215
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 44516 caggtgaagacggacgagcgagtcttgcacacggagcgggggctgctgttccgcaggctt 44457

Query: 2216  agccgtttcgatgcgggcacctacacctgcaccactctggagcatggcttctcccagact 2275
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 44456 agccgtttcgatgcgggcacctacacctgcaccactctggagcatggcttctcccagact 44397

Query: 2276  gtggtccgcctggctctggtggtgattgtggcctcacagctggacaacctgttccctccg 2335
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 44396 gtggtccgcctggctctggtggtgattgtggcctcacagctggacaacctgttccctccg 44337

Query: 2336  gagccaaagccagaggagcccccagcccggggaggcctggcttccaccccacccaaggcc 2395
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 44336 gagccaaagccagaggagcccccagcccggggaggcctggcttccaccccacccaaggcc 44277

Query: 2396  tggtacaaggacatcctgcagctcattggcttcgccaacctgccccgggtggatgagtac 2455
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 44276 tggtacaaggacatcctgcagctcattggcttcgccaacctgccccgggtggatgagtac 44217

Query: 2456  tgtgagcgcgtgtggtgcaggggcaccacggaatgctcaggctgcttccggagccggagc 2515
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 44216 tgtgagcgcgtgtggtgcaggggcaccacggaatgctcaggctgcttccggagccggagc 44157
```

```
Query: 2516  cggggcaagcaggccaggggcaagagctgggcagggctggagctaggcaagaagatgaag 2575
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 44156 cggggcaagcaggccaggggcaagagctgggcagggctggagctaggcaagaagatgaag 44097


Query: 2576  agccgggtgcatgccgagcacaatcggacgccccgggaggtggaggccacgtag 2629
              ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 44096 agccgggtgcatgccgagcacaatcggacgccccgggaggtggaggccacgtag 44043


 Score =  781 bits (394), Expect = 0.0
 Identities = 394/394 (100%)
 Strand = Plus / Minus

Query: 2     atggcctgtgccctagctgggaaggtcttcccaatggggagctggccagtgtggcacaaa 61
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 53746 atggcctgtgccctagctgggaaggtcttcccaatggggagctggccagtgtggcacaaa 53687


Query: 62    agcctgcactgggccaacaaggtggaaggagaagcggcaggtggacggcaaggccccagc 121
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 53686 agcctgcactgggccaacaaggtggaaggagaagcggcaggtggacggcaaggccccagc 53627


Query: 122   ctccttctctcctccgcccctcttcccgcccaggactgggtggagccactgccttataag 181
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 53626 ctccttctctcctccgcccctcttcccgcccaggactgggtggagccactgccttataag 53567


Query: 182   tggtggcctggtggcagcagagcaaactacaaccggcggccagcgggaccagagggcggc 241
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 53566 tggtggcctggtggcagcagagcaaactacaaccggcggccagcgggaccagagggcggc 53507


Query: 242   tctgcaggcaggcggcagcggtgccctcagttccccagcatggccccctcggcctgggcc 301
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 53506 tctgcaggcaggcggcagcggtgccctcagttccccagcatggccccctcggcctgggcc 53447


Query: 302   atttgctggctgctaggggcctcctgctccatggggtagctctggccccagccccggc 361
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 53446 atttgctggctgctaggggcctcctgctccatggggtagctctggccccagccccggc 53387


Query: 362   cccagtgtgccccgcctgcggctctcctaccgag 395
              ||||||||||||||||||||||||||||||||
Sbjct: 53386 cccagtgtgccccgcctgcggctctcctaccgag 53353


 Score =  462 bits (233), Expect = e-127
 Identities = 233/233 (100%)
 Strand = Plus / Minus

Query: 1423  gtgccccagcaagatgaccgcacagccaggacggccttttggcagcaccaaggactaccc 1482
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
```

```
Sbjct: 48539 gtgccccagcaagatgaccgcacagccaggacggccttttggcagcaccaaggactaccc 48480


Query: 1483  agatgaggtgctgcagtttgcccgagcccacccccctcatgttctggcctgtgcggcctcg 1542
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 48479 agatgaggtgctgcagtttgcccgagcccacccccctcatgttctggcctgtgcggcctcg 48420


Query: 1543  acatggccgccctgtccttgtcaagacccacctggcccagcagctacaccagatcgtggt 1602
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 48419 acatggccgccctgtccttgtcaagacccacctggcccagcagctacaccagatcgtggt 48360


Query: 1603  ggaccgcgtggaggcagaggatgggacctacgatgtcattttcctggggactg 1655
              |||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 48359 ggaccgcgtggaggcagaggatgggacctacgatgtcattttcctggggactg 48307


 Score =  456 bits (230), Expect = e-125
 Identities = 230/230 (100%)
 Strand = Plus / Minus


Query: 1789  gcaaatgctatacgtgggctctcggctgggtgtggcccagctgcggctgcaccaatgtga 1848
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 46640 gcaaatgctatacgtgggctctcggctgggtgtggcccagctgcggctgcaccaatgtga 46581


Query: 1849  gacttacggcactgcctgtgcagagtgctgcctggcccgggacccatactgtgcctggga 1908
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 46580 gacttacggcactgcctgtgcagagtgctgcctggcccgggacccatactgtgcctggga 46521


Query: 1909  tggtgcctcctgtacccactaccgccccagccttggcaagcgccggttccgccggcagga 1968
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 46520 tggtgcctcctgtacccactaccgccccagccttggcaagcgccggttccgccggcagga 46461


Query: 1969  catccggcacggcaaccctgccctgcagtgcctgggccagagccaggaag 2018
              ||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 46460 catccggcacggcaaccctgccctgcagtgcctgggccagagccaggaag 46411


 Score =  327 bits (165), Expect = 2e-86
 Identities = 165/165 (100%)
 Strand = Plus / Minus


Query: 394   agacctcctgtctgccaaccgctctgccatctttctgggcccccagggctccctgaacct 453
              |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 51349 agacctcctgtctgccaaccgctctgccatctttctgggcccccagggctccctgaacct 51290


Query: 454   ccaggccatgtacctagatgagtaccgagaccgcctctttctgggtggcctggacgccct 513
              |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 51289 ccaggccatgtacctagatgagtaccgagaccgcctctttctgggtggcctggacgccct 51230
```

```
Query: 514   ctactctctgcggctggaccaggcatggccagatccccgggaggt 558
             ||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 51229 ctactctctgcggctggaccaggcatggccagatccccgggaggt 51185


 Score =  294 bits (148), Expect = 3e-76
 Identities = 148/148 (100%)
 Strand = Plus / Minus

Query: 1276  cagtgccgtgttccagggcttcgccgtctgtgtgtaccacatggcagacatctgggaggt 1335
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 48964 cagtgccgtgttccagggcttcgccgtctgtgtgtaccacatggcagacatctgggaggt 48905

Query: 1336  tttcaacgggccctttgcccaccgagatgggcctcagcaccagtgggggccctatggggg 1395
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 48904 tttcaacgggccctttgcccaccgagatgggcctcagcaccagtgggggccctatggggg 48845

Query: 1396  caaggtgcccttccctcgccctggcgtg 1423
             |||||||||||||||||||||||||||
Sbjct: 48844 caaggtgcccttccctcgccctggcgtg 48817


 Score =  292 bits (147), Expect = 1e-75
 Identities = 147/147 (100%)
 Strand = Plus / Minus

Query: 947   gacccccggtttgtgatggccgcccggatccctgagaactctgaccaggacaatgacaag 1006
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 49850 gacccccggtttgtgatggccgcccggatccctgagaactctgaccaggacaatgacaag 49791

Query: 1007  gtgtacttcttcttctcggagacggtcccctcgcccgatggtggctcgaaccatgtcact 1066
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 49790 gtgtacttcttcttctcggagacggtcccctcgcccgatggtggctcgaaccatgtcact 49731

Query: 1067  gtcagccgcgtgggccgcgtctgcgtg 1093
             |||||||||||||||||||||||||||
Sbjct: 49730 gtcagccgcgtgggccgcgtctgcgtg 49704


 Score =  286 bits (144), Expect = 6e-74
 Identities = 145/146 (99%)
 Strand = Plus / Minus

Query: 2017  agaagaggcagtgggacttgtggcagccaccatggtctacggcacggagcacaatagcac 2076
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 46108 agaagaggcagtgggacttgtggcagccaccatggtctacggcacggagcacaatagcac 46049

Query: 2077  cttcctggagtgcctgcccaagtctccccargctgctgtgcgctggctcttgcagaggcc 2136
             |||||||||||||||||||||||||||| ||||||||||||||||||||||||||||
```

```
Sbjct: 46048 cttcctggagtgcctgcccaagtctccccaggctgctgtgcgctggctcttgcagaggcc 45989


Query: 2137  aggggatgaggggcctgaccaggtga 2162
              |||||||||||||||||||||||||
Sbjct: 45988 aggggatgaggggcctgaccaggtga 45963


 Score =  240 bits (121), Expect = 3e-60
 Identities = 121/121 (100%)
 Strand = Plus / Minus


Query: 619   gacagagtgcgccaacttcgtgcgggtgctacagcctcacaaccggacccacctgctagc 678
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 50745 gacagagtgcgccaacttcgtgcgggtgctacagcctcacaaccggacccacctgctagc 50686


Query: 679   ctgtggcactggggccttccagcccacctgtgccctcatcacagttggccaccgtgggga 738
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 50685 ctgtggcactggggccttccagcccacctgtgccctcatcacagttggccaccgtgggga 50626


Query: 739   g 739
              |
Sbjct: 50625 g 50625


 Score =  236 bits (119), Expect = 5e-59
 Identities = 119/119 (100%)
 Strand = Plus / Minus


Query: 829   agacggggagctgtacacgggtctcactgctgacttcctggggcgagaggccatgatctt 888
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 50132 agacggggagctgtacacgggtctcactgctgacttcctggggcgagaggccatgatctt 50073


Query: 889   ccgaagtggaggtcctcggccagctctgcgttccgactctgaccagagtctcttgcacg 947
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 50072 ccgaagtggaggtcctcggccagctctgcgttccgactctgaccagagtctcttgcacg 50014


 Score =  230 bits (116), Expect = 3e-57
 Identities = 116/116 (100%)
 Strand = Plus / Minus


Query: 1093  gaatgatgctgggggccagcgggtgctggtgaacaaatggagcactttcctcaaggccag 1152
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 49489 gaatgatgctgggggccagcgggtgctggtgaacaaatggagcactttcctcaaggccag 49430


Query: 1153  gctggtctgctcggtgcccggccctggtggtgccgagacccactttgaccagctag 1208
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 49429 gctggtctgctcggtgcccggccctggtggtgccgagacccactttgaccagctag 49374
```

```
Score =  188 bits (95), Expect = 1e-44
Identities = 95/95 (100%)
Strand = Plus / Minus

Query: 1655  gactcagggtctgtgctcaaagtcatcgctctccaggcagggggctcagctgaacctgag 1714
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 48212 gactcagggtctgtgctcaaagtcatcgctctccaggcagggggctcagctgaacctgag 48153

Query: 1715  gaagtggttctggaggagctccaggtgtttaaggt 1749
             |||||||||||||||||||||||||||||||||||
Sbjct: 48152 gaagtggttctggaggagctccaggtgtttaaggt 48118


Score =  184 bits (93), Expect = 2e-43
Identities = 93/93 (100%)
Strand = Plus / Minus

Query: 738   agcatgtgctccacctggagcctggcagtgtggaaagtggccgggggcggtgccctcacg 797
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 50351 agcatgtgctccacctggagcctggcagtgtggaaagtggccgggggcggtgccctcacg 50292

Query: 798   agcccagccgtccctttgccagcaccttcatag 830
             |||||||||||||||||||||||||||||||||
Sbjct: 50291 agcccagccgtccctttgccagcaccttcatag 50259


Score =  143 bits (72), Expect = 6e-31
Identities = 72/72 (100%)
Strand = Plus / Minus

Query: 1207  agaggatgtgttcctgctgtggcccaaggccgggaagagcctcgaggtgtacgcgctgtt 1266
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 49265 agaggatgtgttcctgctgtggcccaaggccgggaagagcctcgaggtgtacgcgctgtt 49206

Query: 1267  cagcaccgtcag 1278
             ||||||||||||
Sbjct: 49205 cagcaccgtcag 49194


Score =  129 bits (65), Expect = 9e-27
Identities = 65/65 (100%)
Strand = Plus / Minus

Query: 555   aggtcctgtggccaccgcagccaggacagagggaggagtgtgttcgaaagggaagagatc 614
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 51063 aggtcctgtggccaccgcagccaggacagagggaggagtgtgttcgaaagggaagagatc 51004
```

```
Query: 615   ctttg 619
             |||||
Sbjct: 51003 ctttg 50999
```

Score = 87.8 bits (44), Expect = 3e-14
Identities = 44/44 (100%)
Strand = Plus / Minus

```
Query: 1746  aggtgccaacacctatcaccgaaatggagatctctgtcaaaagg 1789
              |||||||||||||||||||||||||||||||||||||||||||
Sbjct: 47403 aggtgccaacacctatcaccgaaatggagatctctgtcaaaagg 47360
```

>AC000063.1.1.34478
Length = 34478

Score = 71.9 bits (36), Expect = 2e-09
Identities = 48/52 (92%)
Strand = Plus / Minus

```
Query: 1860 ctgcctgtgcagagtgctgcctggcccgggacccatactgtgcctgggatgg 1911
              |||||||||| || ||||||| |||||||||| ||||||||||||||||
Sbjct: 5711 ctgcctgtgctgactgctgccttgcccgggacccttactgtgcctgggatgg 5660
```

>AC079799.7.1.172495
Length = 172495

Score = 54.0 bits (27), Expect = 5e-04
Identities = 42/47 (89%)
Strand = Plus / Minus

```
Query: 1865   tgtgcagagtgctgcctggcccgggacccatactgtgcctgggatgg 1911
               ||||| || |||||||||||| || ||||| ||||||||||||||||
Sbjct: 151014 tgtgctgactgctgcctggctcgagacccttactgtgcctgggatgg 150968
```

Database: Homo_sapiens.latestgp.fa
Posted date: Jul 8, 2003 12:51 PM
Number of letters in database: 200,800,637,119
Number of sequences in database: 26,679

Lambda	K	H
1.37	0.711	1.31

Gapped

Lambda	K	H
1.37	0.711	1.31

Matrix: blastn matrix:1 -3
Gap Penalties: Existence: 0, Extension: 0

```
Number of Hits to DB: 0
length of query: 5260
length of database: 200,800,637,119
effective HSP length: 22
effective length of query: 2607
effective search space used:          0
T: 0
A: 0
X1: 0 ( 0.0 bits)
X2: 20 (39.7 bits)
S1: 12 (24.3 bits)
S2: 24 (48.1 bits)
```